THE DIVERSITY AND ECOLOGY OF NITROGEN-FIXING BACTERIA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

John Christian Gaby

May 2013

THE DIVERSITY AND ECOLOGY OF NITROGEN-FIXING BACTERIA

John Christian Gaby, Ph. D.

Cornell University 2013

Biological nitrogen fixation, the enzymatic conversion of gaseous nitrogen to ammonium, is carried out by diverse *Bacteria* and *Archaea*. Neither the diversity of nitrogen fixers nor the factors which control nitrogen fixation in the environment are well-characterized. The *nifH* gene encodes a component of the enzyme which carries out nitrogen fixation. I describe a *nifH* database, its sequence composition, and its inherent limitations. As an application, the database was used in an evolutionary analysis which compared the *nifHDK* genes and the 16S rRNA gene, finding support for the horizontal transfer of nitrogen fixation between the *Bacteria* and *Archaea*. Next, the database was used to evaluate the global diversity of nitrogen fixers. At an $OTU_{0.95}$ similarity cutoff, 3,358 *nifH* OTUs were identified. Soil was found to be contain more *nifH* diversity than marine waters, and cluster III had the highest richness in the *nifH* phylogeny. The database was next used to evaluate the *in silico* coverage of more than 50 published, universal *nifH* primers, to find that 23 primers target <50% of the known *nifH* diversity while 15 target >90%. The set DVV/IGK3 had the best laboratory performance against genomic and soil DNA. Next, 5 full-length, *nifH* gene templates representing the full diversity of nitrogen-fixers were used against degenerate, *nifH* primer sets to assess the potential for template-specific bias in quantitative real-time PCR. It was found that bias can lead to over 3 orders of magnitude mis-estimation of template copy number, with the degree of bias dependent on the compared templates, the primer set, and the annealing temperature used. The

primer set polF/polR exhibited the least template-specific bias. This primer set was used in an ecological survey of 20 old-field and meadow sites from the Finger Lakes and Champlain regions of New York State. Either nitrogen fixation or nitrogen mineralization tended to dominate in a site, and this corresponded to the $\delta\ ^{13}C$ signature of the soil, suggesting that the processes may be driven by distinct biomass inputs. Nitrogen fixation was found to be constrained in most sites, and soil moisture was identified to constrain the abundance of nitrogen fixers.

BIOGRAPHICAL SKETCH

Chris was born and raised in the rolling foothills of the mountains of East Tennessee where he lived until he completed his undergraduate studies in 2002 and received a B.S. degree in Biology with emphasis in Microbiology. In December 2002, he took his first ever flight outside his home country and began 27 months of Peace Corps service in the West African country Niger where he learned the local language Zarma, acclimated to the culture, and carried out agricultural development projects. After returning to the U.S., he began a doctoral program in Microbiology and joined the laboratory of Dr. Daniel Buckley in the Cornell Crop and Soil Sciences Department. In the Buckley laboratory his research focused on nitrogen fixation and the use of the marker gene *nifH* as a tool to study nitrogen-fixing bacteria. His work began with a series of bioinformatic approaches to the analysis of an aligned *nifH* database. He later focused on the use of quantitative, real-time PCR as a tool to study the ecology of nitrogen fixers. In the final year of his graduate studies, he was awarded a Fulbright U.S. Student fellowship and spent 11 months in Colombia, where he focused on another aspect of the nitrogen cycle, denitrification, in association with Dr. Maria Mercedes Zambrano of Corporacion Corpogen and with Dr. Esperanza Torres of the Agronomy Department of the National University of Colombia, both located in Bogota. Chris hopes to apply the knowledge and skills he has acquired to carry out research on soil fertility and plant-microbe interactions in order to improve crop productivity and management practices for agricultural development.

ACKNOWLEDGMENTS

I would like to like to thank my doctoral advisor, Dr. Daniel Buckley, for the time, effort, support, and good guidance he has given me over my years as a graduate student. I also thank my minor advisors, Dr. Stephen Zinder, and Dr. Peter Groffman for their guidance.

I thank the staff of the USDA ARS Culture Collection for their service providing strains used in this work. I also thank those individuals who provided strains or DNAs including Dr. Esther Angert and Lilly Bojarski for sharing various cultures, Dr. James Shapleigh for providing a *Rhodobacter sphaeroides* culture, Dr. Eugene Madsen and Buck Hanson for providing *Polaromonas naphthalenivorans* genomic DNA, Eric G Luning, Dr. Aindrila Mukhopadhyay, and Dr. Jay D Keasling (Lawrence Berkeley National Laboratory) for providing *Desulfovibrio vulgaris* genomic DNA, Dr. Derek Lovley and Joy Ward for providing a *Geobacter uraniireducens* culture, and Dr. David Benson and Ying Huang for providing *Frankia* sp. genomic DNA.

I thank Dr. Christopher Jones for advice using the software ARB.

I thank Dr. Sean Berthrong and Ashley Campbell for comments on a draft manuscript.

I thank those individuals who facilitated sampling including Amy Ivy of the Clinton County Cornell Cooperative Extension, The USDA Forest Service and the staff of the Fingerlakes National Forest including Christopher Zimmer, Michael C. Liu, and Nancy Burt, Karl Pendleton of Cornell Caldwell Farm, Gary Tenant of Cornell Mount Pleasant Farm, Dale Cotterill of the Cornell Aurora Musgrave Farm,

and Peter Smallidge, Stephen Morreale and Brian Chabot of the Cornell Arnot Forest, Todd Bittner of the Cornell Plantations, Josef Wetzstein of the USDA NRCS, Michael Davis of the Cornell Willsboro Farm, Chris Nobles of the Uihlein Potato Farm, and Michael Farrell of the Uihlein Sugar Bush.

I thank Robert Schindelbeck of the Cornell Soil Health Program who provided advice on soil analysis approaches.

I thank Carolyn Mead for help sampling the Champlain old field and meadow sites.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

As our world grows more populated and demand increases for food, fiber, and biofuels, the need for less energy-intensive, more environmentally-friendly agricultural practices becomes apparent. Biological nitrogen fixation gives hope for a solution that would reduce the need for energy-intensive nitrogen fertilizers. However, despite many decades of work, neither the full scope of the diversity of nitrogen fixers nor the ecological constraints on nitrogen fixation in the environment are well understood. This dissertation is an effort to address our lack of understanding on the diversity and ecology of nitrogen fixers.

The format of the dissertation follows the paper style. This first chapter serves as an introductory chapter, and the final chapter as a summary of further steps that may be taken to expand upon the work. At the time of publication, two of the five main chapters (Chapter 3 and Chapter 4) had already been published while the remaining three main chapters are intended for publication.

Chapter 2

This chapter describes the content and potential biases associated with the *nifH* database. The database described is the most recent update of the database from May 16, 2012, and the databases used for analyses in Chapter 3 and Chapter 4 are earlier versions. The database is a tool for researchers wishing to study the diversity, evolution, or phylogeny of nitrogen-fixing bacteria and may also, as shown in Chapter 4, be used as a first step in evaluating the effectiveness of primers through coverage

analysis. We demonstrate one application of the database by examining the evolutionary history of nitrogen fixation through an analysis of evolutionary distances. Here we compare the nitrogenase structural genes *nifH, nifD,* and *nifK* both to themselves and to the 16S rRNA encoding gene, finding support for horizontal transfer of nitrogen fixation between the *Bacteria* and *Archaea*. This chapter serves as the natural starting point of the dissertation given that the *nifH* database formed a basis for the remaining work contained herein. At the time of submission of the dissertation, the aim is to submit this chapter as a paper to the journal Database.

Chapter 3

In this chapter I detail the results of a global estimate of nitrogen fixer richness as determined by Chao Richness Estimation of the total number of species in the *nifH* database according to an $OTU_{0.95}$ similarity cutoff. In addition to a total estimate of nitrogen fixer diversity, evaluations of environmental and phylogenetic groups identified the terrestrial environment to have greater richness over marine and microbial mat environments while with respect to phylogeny cluster III had the greatest richness among the phylogenetic comparisons. We also established the top 10 most abundant OTUs in the database to find that only half had cultivated representatives and that the OTUs were derived from the *Cyanobacteria* and the *Proteobacteria.* This work suggests environments and phylogenetic groups where there remains considerable diversity to be sampled, and it has been published in the journal Environmental Microbiology (Gaby, J.C. and Buckley, D.H. [2011]. A global census of nitrogenase diversity. 13[7]: 1790-9).

Chapter 4

For this chapter I have compiled a comprehensive list of *nifH* primers, and I describe a combination of bioinformatics and laboratory work that I used to identify the best-performing universal *nifH* primer sets. The primers evaluated were all taken from previously published studies. A map of the binding site of the primers relative to the *nifH* gene shows that many primers overlap in just a few regions which are responsible for the enzyme's biochemical activity. I show that many of the more than 50 universal *nifH* primer sets exhibit poor coverage as determined by evaluating the primers for matches to the sequences in the *nifH* sequence database. Bias was evaluated by assessing the primer coverage within the *nifH* phylogeny. Group-specific *nifH* primers were also identified in the literature and then evaluated in a similar manner to the universal evaluations. Finally, I did an in-laboratory evaluation of universal *nifH* primer performance for those sets that showed a 90% or greater coverage *in silico* whereby I identified the best-performing primer combinations against a set of genomic DNAs from diverse nitrogen fixers as well as with soil DNA. This work has been published in the journal PLoS One (Gaby J.C., and Buckley D.H. [2012] A Comprehensive Evaluation of PCR Primers to Amplify the *nifH* Gene of Nitrogenase. [7]: e42149. doi:10.1371/journal.pone.0042149).

Chapter 5

This work demonstrates the occurrence of template-specific bias exhibited by degenerate *nifH* primers when used in qPCR. By using as standards the full-length, PCR-amplifed *nifH* genes from the genomes of five nitrogen fixers which represent the full breadth of the *nifH* phylogeny, a discrepancy in determined copy number is

shown across *nifH* primer sets despite using standards diluted to identical copy number. The polF/polR primer set exhibited the least copy number variation between the assessed templates. All templates varied in nucleotide composition in the region of primer binding for the primer sets but in no case did they exhibit mismatches to the evaluated primers. In some cases the difference in Ct value would amount to a greater than 1000-fold mis-estimation of copy number depending on the template comparison. Thus, depending on the community composition of the nitrogen fixers present in a sample and the standard used as comparison, a serious over- or under-estimation may occur. At the time of submission of the dissertation, the aim is to submit this as a paper to the journal Applied and Environmental Microbiology.

Chapter 6

This capstone work is an ecological survey of nitrogen fixer abundance in old field sites in two regions of New York State, the Finger Lakes region and the Champlain region. The polF/polR primer set, which was identified in Chapter 5 to exhibit low template-specific amplification bias, was applied in quantitative, real-time PCR assays of *nifH* abundance in soils of 20 old field sites in the two regions. A number of abiotic soil variables were measured along with the nitrogen mineralization rate, and these served as independent variables in statistical analyses seeking to explain two dependent variables: the nitrogen fixation rate measured via stable isotope incorporation and the *nifH* copy number measured via qPCR. While nitrogen fixation rate showed no correlation with independent variables or *nifH* copy number, the *nifH* copy numbers showed several interesting correlations which were affected by the type of normalization. Copy number was first normalized to soil dry weight where a

significant positive linear relationship was observed with the likely inter-related variables Ca, Mg, and pH as well as with soil moisture. Two distinct groups of samples were identified according the isotopic data for C and N, and these groups were compared for soil variables, whereby nitrogen fixation was found to predominate in an isotopically depleted group whereas nitrogen mineralization predominated in the enriched group. For *nifH* copies normalized to the amount of DNA extracted from the soil, *nifH* copies were greater in the depleted group. This work shows that high rates of nitrogen mineralization and nitrogen fixation do not occur in the same site, and that nitrogen fixation is constrained by factors like low soil moisture and biomass inputs.

Chapter 7

This chapter describes future directions that may be taken to follow up on the work described in chapters 2 to 6.

CHAPTER 2

A COMPREHENSIVE, ALIGNED *NIFH* GENE DATABASE: A MULTI-PURPOSE TOOL FOR STUDIES OF NITROGEN-FIXING BACTERIA

Introduction

Biological nitrogen fixation contributes around half of annual nitrogen inputs into the biosphere [1] and is an important source of nitrogen for agriculture [2, 3] and many ecosystems [4]. Biological nitrogen fixation is catalyzed by the nitrogenase enzyme [5] which is found only in the domains *Bacteria* and *Archaea* [6]. Nitrogenase consists of a heterotetrameric core, composed of NifD and NifK, and a dinitrogenase reductase subunit, consisting of NifH. Dinitrogenase reductase transfers reducing equivalents to the core enzyme complex where those reducing equivalents are used to convert N_2 into NH_3 equivalents. The core enzyme contains a Mo-Fe metal cluster whose formation requires NifEN, which are structural homologs of NifDK [7]. These genes are generally found in the order *nifHDKEN,* though in some cases a few intervening genes are observed and on occasion *nifHDK* and *nifEN* will be found in distant locations on the genome [8].

The *nifH* gene is the most widely used biomarker for studying the ecology and

evolution of nitrogen-fixing bacteria in the environment and is also commonly used to examine the evolutionary origins of nitrogen fixation [8]. Surveys of *nifH* diversity have been conducted in a wide range of marine (e.g. [9-11]) and terrestrial environments (e.g.[12-14]). Other studies have explored the diversity of nitrogen-fixers in extreme environments [15] and associated with world commodity crops [12]. Many of these studies seek to characterize the ecology of nitrogen-fixing organisms. The diversity of nitrogen-fixing organisms varies dramatically across habitats, and different habitats select for different types of nitrogen fixers [16, 17]. Furthermore, *nifH* diversity has been associated with rates of nitrogen fixation [18].

Homologs of the *nifH* gene can be divided into five main phylogenetic clusters [19]. Cluster I contains a diverse group of *nifH* genes primarily from aerobic and facultatively anaerobic organisms that belong to phyla including *Proteobacteria*, *Cyanobacteria*, *Firmicutes*, and *Actinobacteria*. Cluster III contains *nifH* genes that are almost exclusively found in obligate anaerobes including methanogenic *Archaea*, sulfate and sulfur reducers, *Treponema* and clostridia. Cluster II contains *anfH*, alternative nitrogenases which are paralogs of *nifH*, and use an Fe-Fe cofactor in place of the Fe-Mo used by *nifH* [20]. There also exist V-Fe alternative nitrogenases which make use of a *nifH* paralog which clusters with *nifH* for reduction of the alternative nitrogenase core. The alternative nitrogenases appear to be found only in the genomes of organisms that also contain *nif* genes [20]. Cluster IV contains further paralogs of *nifH* that in methanogens have a function other than nitrogen fixation [21, 22]. Finally,

a fifth cluster of *nifH*-like genes has been described which contains distant paralogs that function in bacteriochlorophyll synthesis [8, 23, 24]. The presence of paralogous sequences is of concern when the *nif* structural genes are used as markers in environmental surveys of nitrogenase diversity because paralogous sequences may be mistakenly included in *nifH* diversity calculations leading to inflated diversity estimates. However, the vast majority of these paralogs form discrete phylogenetic clusters that are readily distinguishable from *nifH.* The availability of a well curated database makes it easy to identify and remove these sequences prior to further analysis.

Analyses of microbial gene sequences that serve as functional biomarkers are facilitated by the availability of well-characterized databases of aligned sequences. Aligned sequences are required for phylogenetic analyses (e.g. [8, 25-27]), diversity analyses (e.g. [17]), and the design and evaluation of PCR primers (e.g. [28, 29]). For primer and probe design, researchers need a comprehensive gene database that represents existing gene sequence diversity in order to ensure that primers/probes are specific only to the group of interest and that they do not target other groups or closely-related paralogs. Universal and group-specific primers are increasingly being used in quantitative real-time PCR for the quantification of gene copy numbers in the environment (e.g. [30-32]), for expression studies [33], or for tracking organisms (e.g. [30, 34]). Thus, many of the contemporary techniques in molecular microbial ecology are based upon the use of PCR with specific primers which must be designed and

evaluated using sequence databases.

There is strong demand for the creation of a centralized, well-described, aligned and vetted *nifH* database. Similar databases of 16S rRNA gene sequences have been of great utility in the field of microbial ecology [35-37]. In this work, we describe a comprehensive *nifH* database, a multi-purpose tool to facilitate the work of researchers who use sequence-based techniques to study nitrogen-fixing bacteria. The database was assembled using the ARB software package [38] which allows for the compilation and association of aligned sequence data, sequence metadata, and phylogenetic trees. Phylogenetic trees can be used to navigate the sequences and to explore phylogenetic patterns found in associated metadata. Though this *nifH* database has not been described previously, preliminary versions of this *nifH* database, one containing 16,989 and the other 23,843 sequences, were used to evaluate the diversity of *nifH* genes in different environments [17] and to evaluate PCR primers used in environmental surveys of *nifH* diversity [28]. Here we release the completed *nifH* database containing 32,954 aligned *nifH* sequences along with guide trees. We describe trends in sequence acquisition and examine the characteristics of the *nifH* database to raise awareness of the potential limitations and biases associated with the sequence data currently available. The new release also incorporates *nifD*, *nifK*, and 16S rRNA genes from 185 sequenced genomes. These new data layers facilitate improved evolutionary analysis and enable assessment of ancestral patterns of horizontal gene exchange. The database and its associated description will be useful

for studying the evolution, phylogeny, and diversity of nitrogen-fixing bacteria.

Materials and Methods

Database construction

The database comprises all *nifH* sequences available in the Genbank nucleotide database as of May 16, 2012. The majority of sequences are derived from environmental surveys of *nifH* diversity, though the database also contains sequences obtained from a range of isolates. Sequence alignment was performed using the ARB integrated aligner employed through an iterative process to generate a nucleotide alignment consistent with the Pfam [39] Fer4_NifH amino acid alignment. We first downloaded the Pfam Fer4_NifH (PF00142) alignment for use as an alignment template and reverse-translated this alignment using BioEdit [40]. This reverse-translated alignment was imported into ARB and used as a template to generate a seed alignment consisting of a wide phylogenetic diversity of full length and near full length *nifH* nucleotide sequences from Genbank. The reverse-translated sequences were then deleted from the database, the seed alignment was manually vetted, and then this seed alignment was used to complete the alignment of all remaining sequences with the ARB integrated aligner. The final alignment was vetted manually, and low quality sequences were removed.

Using data from sequenced genomes, records for *nifD*, *nifK*, and 16S rRNA genes were added to the database and linked to their corresponding *nifH* entries. The NifD and NifK sequences were aligned independently using Clustal [41]. The aligned sequences were imported into ARB as separate DNA sequence alignments and merged with the existing *nifH* records. For import of the 16S rRNA gene sequences, we used the Greengenes [42] alignment, which allowed us to import the 16S rRNA helix data from the Greengenes ARB database. In some cases 16S rRNA gene sequences for the queried genomes were not present in the Greengenes database. In these cases 16S rRNA gene sequences were extracted from the Genbank genome records and the NAST aligner was used to align these sequences in Greengenes. All the aligned 16S rRNA gene sequences were imported into the ARB *nifH* database and merged with the *nifHDK* genes from their respective genomes.

Analysis

Analyses were performed to evaluate the sequence entries present in the *nifH* database. The number of *nifH* sequence entries was compared to the number of entries for other important functional genes used in microbial ecology by searching the Genbank nucleotide database. The number of *nifH* sequences contributed to Genbank over time was determined by a search of the Genbank nucleotide database for the term *nifH* in the gene field of the records for the given years. To obtain the number of

sequences contributed per study, we examined the title field of each sequence to determine the number of unique titles. This approach assumes that each study has only one title associated with the sequences submitted for the study, which we generally observed to be true.

Sequence characteristics were calculated in ARB. Sequence lengths and percent G+C content were calculated using the ARB Node Display Setup (NDS). The percent G+C contents for *nifH*, *nifD*, and *nifK* were compared to the average percent G+C for the genome from which the genes were derived. The percent genome G+C content was obtained from the table of prokaryotic genomes available at NCBI (http://www.ncbi.nlm.nih.gov/genome/browse/). The start and end positions of sequences were calculated using the *nifH* gene of *Azotobacter vinlandii* (Genbank ACCN# M20568) as the reference sequence to provide sequence position.

Sequence affiliation with country and environment of origin was determined using the search and query function of ARB. Searches for country and affiliation of countries to a continent (Figure 2.3 A and B) were done according to the United Nations "standard country or area codes and geographical regions for statistical use" [51, 52]. Fields listing environment data like isolation source, title, source, and note were used to determine the number of sequences associated with major environments and environments of special interest (Table 2.1). As confirmation that sequences were affiliated with the correct environment, both the list of marked sequences as well as those not marked for the environmental category were visually vetted. There were

4,759 sequences that did not have an annotated isolation source field. The habitat categories are not mutually exclusive. For example, sequences that fall in categories such as maize or rice will also belong to categories such as soil and agricultural. The category marine refers to marine water column and does not include marine sediments. The category terrestrial includes phyllosphere and rhizosphere regardless of plant type. The category mat includes phototrophic and chemolithotrophic microbial mats from marine and geothermal environments. The category thermal includes hyperthermal vents. The category soil includes rhizosphere and bulk soil. The category termite includes termite gut contents. Sequences belonging to the individual environmental categories are saved as configurations in the ARB *nifH* database for further reference.

Evolutionary distances were calculated using in ARB without distance correction or weighting. Filters were applied to exclude uncertain regions of alignment and indel regions, including a prominent indel in *nifD.* These filters are available within the *nifH* database. Distance matrices for each gene were determined in ARB and exported for analysis in R [45]. Affiliation of individual gene sequences with *nifH* phylogenetic clusters was established via neighbor joining trees that were generated in ARB and stored as part of the *nifH* database. The database includes two types of trees. First is the tree of all *nifH* sequences which was generated by constructing a neighbor joining, base tree of thousands of the longer sequences with overlapping nucleotide positions in the alignment. To this tree the remaining sequences were added using the quick add by parsimony feature of ARB. The second set of trees are neighbor joining

trees of each set of the full-length sequences for the *nif* and 16S genes.

All statistical analyses were performed using the statistical package R [46].

Results

Analysis of sequence-associated metadata

On July 12, 2012 there were 34,160 *nifH* sequence entries in Genbank, making *nifH* one of the most heavily sequenced microbial functional genes (data not shown). Other highly sequenced microbial functional genes include the *rbcL* gene of RUBISCO and the *amoA* gene of ammonia monooxygenase, the first step in nitrification (data not shown). The number of *nifH* genes accumulating each year continues on an upward trend (Figure 2.1) with 7,702 sequence depositions in 2011.

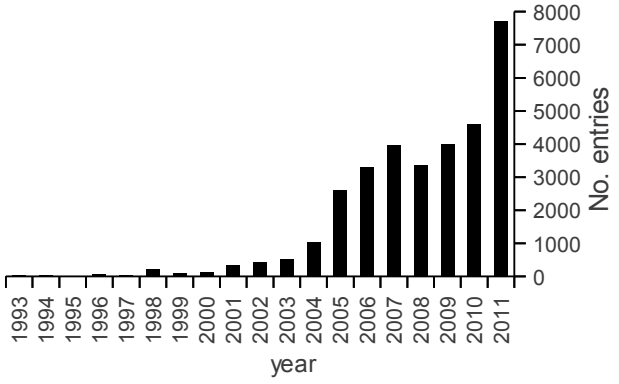


Figure 2.1 The yearly accumulation of *nifH* sequences in the Genbank nucleotide database.

The database consists of 32,954 aligned sequences, the majority of which are *nifH* though we have included paralogs such as alternative nitrogenases and some cluster IV sequences for reference. There are 490 full-length sequences homologous to *nifH* in the database, 375 of these are *nifH* including 245 sequences from sequenced genomes. The full-length *nifH* sequences represent 294 different strains and 222 different species.

There is great variation in the number of sequences representing different

Table 2.1. Habitat categories and the number of sequences associated with each.

Environment/Type	Number of Sequences
terrestrial	17,340
soil	8,568
forest	2,534
grass	416
thermal	225
termite	590
marine	5,958
lake	430
wastewater, sewage	69
Spartina	866
mine spoils	156
mat	5,141
stem	1,407
agricultural	7,308
maize	3,082
rice	854
crust	60
endophytic	1,645

environmental categories (Table 2.1). The most comprehensive environmental categories are terrestrial with 17,340 sequences and marine with 5,958. The majority of *nifH* sequences in the database originate from terrestrial agricultural environments with 7,308 sequences in this category. Numerous studies have focused on plant-microbe associations, many of which examine maize and rice. A total of 3,082 *nifH* sequences have been reported from maize, whereas 854 have been reported from rice. Plant-associated *nifH* sequences are also reported from sugarcane, sorghum, sweet potato, and tomato, as well as sphagnum, Kallar grass, and Douglas fir. Other terrestrial environments that have been sampled for *nifH* sequences include forests with 2,534 sequences in the database, and grasslands with 416. In addition, a total of 978 *nifH* sequences are reported from freshwater environments such as lakes and rivers. Anoxic environments are poorly represented in the database. Major anoxic environments from which *nifH* sequences have been sampled include microbial mats, flooded rice paddy soils, wetlands, sediments, and termite guts.

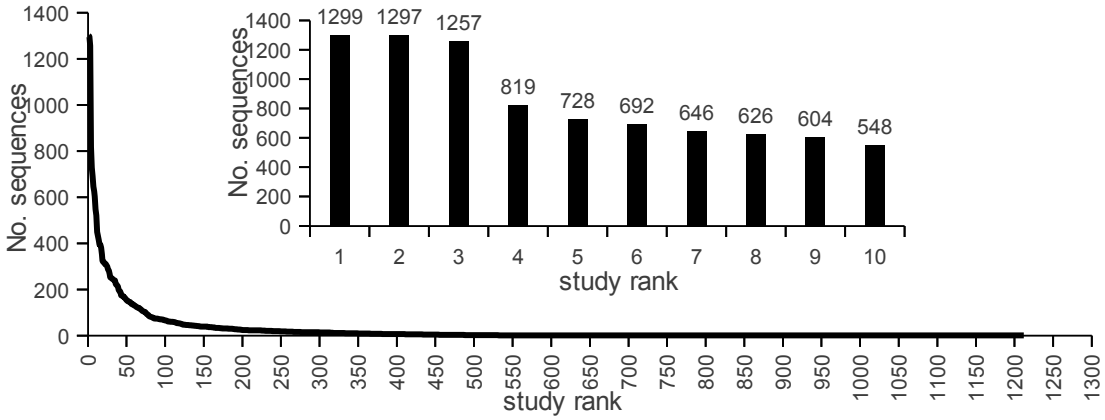


Figure 2.2 The number of sequences associated with each study in the database.

There are 1,211 studies that contributed anywhere from 1 to 1,299 (Figure 2.2) sequences with a mean contribution of 27 sequences per study. There are 584 studies that contributed only 1 sequence (48% of the total number of studies and 1.8% of the sequences in the database) and these represent sequences that originate from isolated strains. The two studies contributing 1,299 and 1,297 sequences originate from two maize studies in Brazil [13], while the study with 1,257 sequences focused on coastal microbial mats [47]. The 10 studies that contributed the most sequences account for 8,516 sequences in total (26% of all sequences in the database) and the top 35 studies contributed 50% of all sequences.

There are 19,445 sequence records (59% of the 32,954 total sequences) that document country of origin. Sampling efforts appear to have neglected locations in Africa and Oceana relative to other regions (Figure 2.3A). Two studies of Brazilian maize **[13]** account for 79% of the sequences from South America. Outside of this single large study only 665 sequences remain to represent all of South America. Of the top 10 countries annotated as sequence sources, the Netherlands falls at the top with 3,502 sequences, but Brazil follows with 3,261 sequences (Figure 2.3B).

Sequence characteristics

The main phylogenetic tree created to facilitate database navigation contains 32,931 sequences. There are 7,030 sequences (21.35% of the total database) in *nifH*

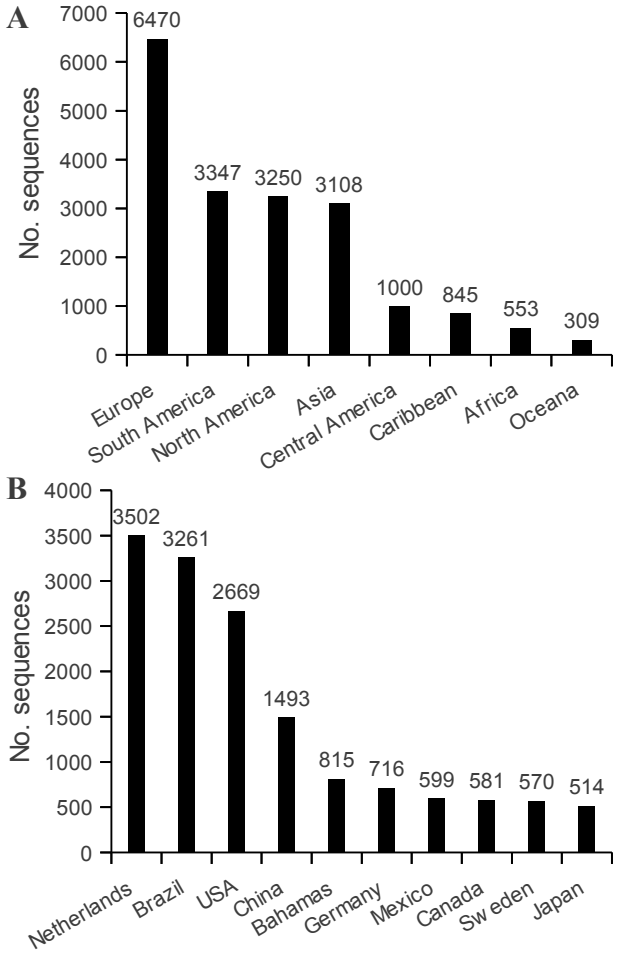


Figure 2.3 (A) Sequences obtained by continent or major geographic region. A total of 9,589 sequences had an annotated country field in the Genbank record out of 17,142 sequences in our *nifH* database. 'Other' consists of Pacific and Atlantic Ocean and Antarctic sequences. (B) The 10 countries with the most *nifH* sequences.

cluster III and subcluster IA, the clusters that contain primarily obligate anaerobes. There are 24,799 sequences (75.31%) in *nifH* cluster I, which contains primarily aerobes and a few facultative anaerobes. The database also contains 578 sequences (1.76%) in *nifH* cluster II which comprises the iron-dependent alternative nitrogenase cluster, and 524 sequences in cluster IV which contains genes not functionally associated with nitrogenase

The mode fragment length for *nifH* gene sequences in the database is 324 bases (Figure 2.4). Environmental surveys of *nifH* employ PCR primers that bind at conserved sites within *nifH* to amplify short sequence fragments. Full-length *nifH* sequences are derived mostly from genome sequences of *Bacteria* and *Archaea* (e.g. [48]) but also include some cloned full-length sequences (e.g. [49, 50]). Different studies have used a diversity of PCR primer sets resulting in a range of overlapping

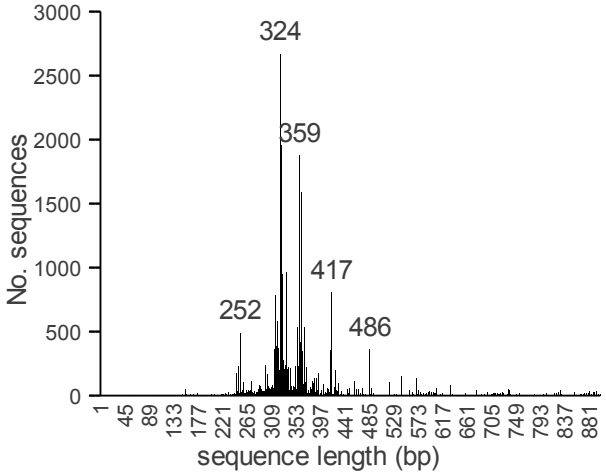


Figure 2.4 Histogram of sequence lengths of the *nifH* database.

sequence fragments. We have identified the frequency of different start and end positions within the database (Figure 2.5). The most common start position for *nifH* primers is at about 100 bases from the start codon, and the most common stop position at about 450 bases from the start codon (Figure 2.5). The distance between the two most abundant start positions as well as stop positions (Figure 2.5) is 18 and 17 bases, which corresponds to the length of a primer suggesting that in some cases primer sequences have not been trimmed from deposited sequences.

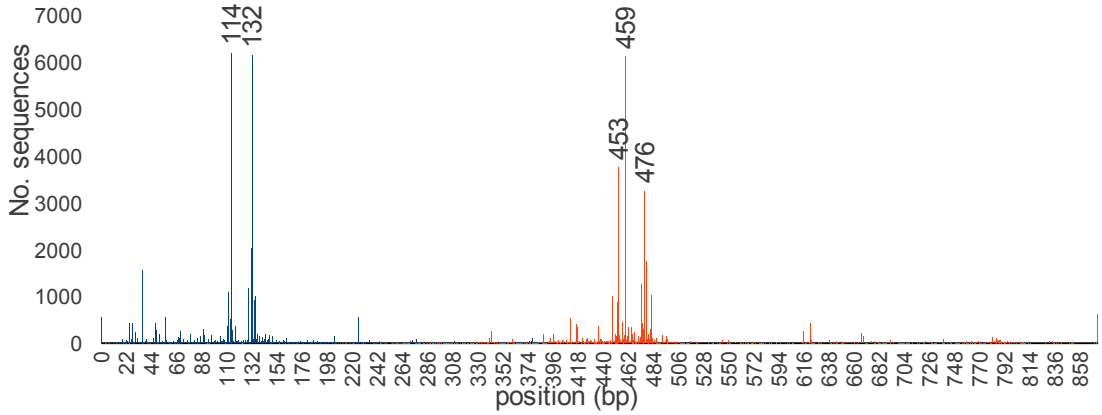


Figure 2.5 Histogram of start (blue) and stop (red) positions for the *nifH* sequences in the database with reference to the positions of the *Azotobacter vinelandii nifH* gene (Genbank ACCN# M20568).

The mean G+C content for all *nifH* gene sequences in the database is 56.8%, however,the G+C content distribution of the *nifH* gene fragments is bimodal (Figure 2.6). The number of sequences in the lower part of the distribution (G+C content of 53% or less) is 9,543 and represent 29% of the *nifH* sequences in the database. The low G+C content sequences are comprised primarily of *Cyanobacteria* (5,157 sequences) and clostridia (887 sequences). Both of these phylogenetic groups are well

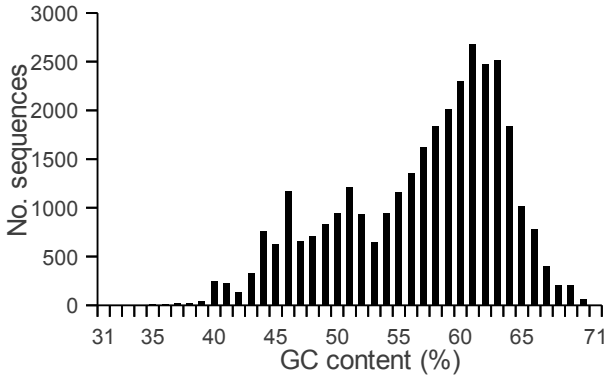


Figure 2.6 Distribution of *nifH* sequences as a function of their percent G+C content.

known to contain organisms with low genome G+C content. The low G+C sequences also include proteobacterial sequences (1,361 sequences), cluster II sequences (330 sequences), and cluster IV sequences (254 sequences).

Comparative analyses of *nifH*, *nifD*, and *nifK* G+C

We identified 185 instances of sequenced genomes that contained *nifH*, *nifD*, *nifK*, and 16S rRNA genes. These multi-gene database entries were used to analyze the G+C content and evolutionary distances for the *nif* genes. There was a linear relationship between genome G+C content and the G+C content of the 3 structural *nif* genes (Figure 2.7; R^2 for *nifH*, *nifD*, and *nifK* was 0.93, 0.92, and 0.90 respectively). The range of genome G+C content was 26.6% to 74.8%. In contrast, *nifH* G+C content ranged from 33.6% to 67.8%, *nifD* from 32.5% to 67.0%, and *nifK* from

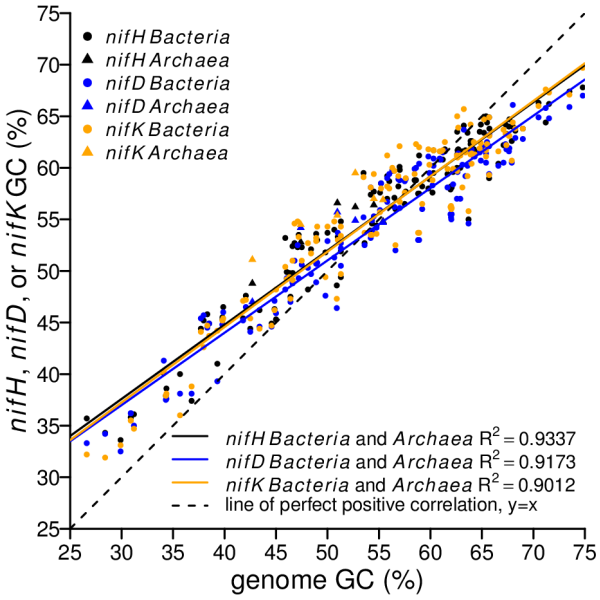
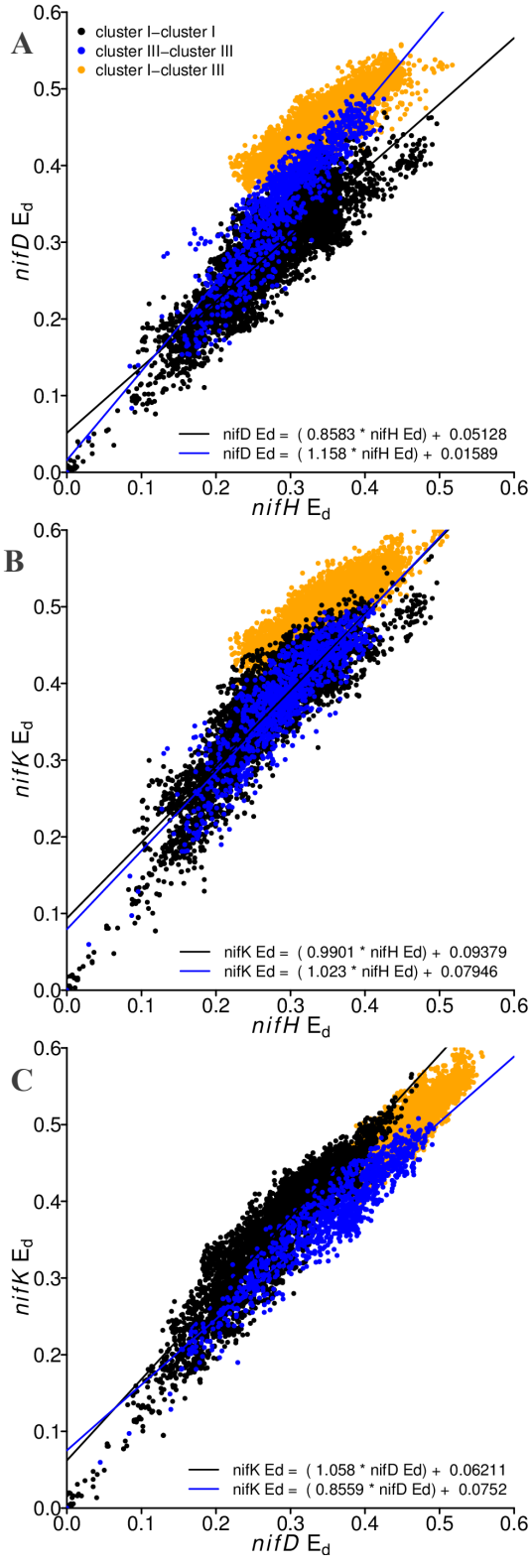


Figure 2.7 Sequence G+C content G+C of complete genomes and corresponding *nifH, nifD,* and *nifK* genes for 6 *Archaea* and 131 *Bacteria*. Values are expressed as percents, colors represent different gene types, and symbols indicate *Bacteria* and *Archaea* as shown in the figure legend. Values are calculated for the full length gene sequences.

31.9% to 69.7%. Outliers indicative of recent horizontal gene transfer were not observed.

Evolutionary distance (Ed) values for the *nifHDK* genes from the 185 genomes were determined (Figure 2.8). Pairwise comparison of Ed values can provide evidence for horizontal transfer or mutation rate variation between genes [51]. The *nifH*, *nifD*, and *nifK* genes display divergent patterns of Ed (Figure 2.8). The sequence divergence between cluster I and III for *nifH* is less than would be expected given the relative rate at which mutations accumulate within clusters for *nifH, nifD,* and *nifK,* and these data suggest the presence of an ancient horizontal gene transfer of *nifH* between cluster I and cluster III diazotrophs (Figure 2.8A and B). In addition, the data indicate mutation

Figure 2.8 Comparison of evolutionary distances (Ed) for *nifH* with *nifD* (A), *nifH* with *nifK* (B), and *nifD* with *nifK* (C). The equation for the regression lines and their coefficient of linear regression (R^2) are labeled in each panel of the figure. Intra-cluster and inter-cluster comparisons are indicated using different colors as defined in the legend. Distances are calculated after exclusion of both indel regions and regions of dubious positional homology Colors indicate intra-domain and inter-domain comparisons for each set of genes as indicated in the figure legend.


A
cluster I–cluster I
cluster III–cluster III
cluster I–cluster III
nifD Ed = (0.8583 * nifH Ed) + 0.05128
nifD Ed = (1.158 * nifH Ed) + 0.01589
nifD Ed
nifH Ed
B
nifK Ed = (0.9901 * nifH Ed) + 0.09379
nifK Ed = (1.023 * nifH Ed) + 0.07946
nifK Ed
nifH Ed
C
nifK Ed = (1.058 * nifD Ed) + 0.06211
nifK Ed = (0.8559 * nifD Ed) + 0.0752
nifK Ed
nifD Ed

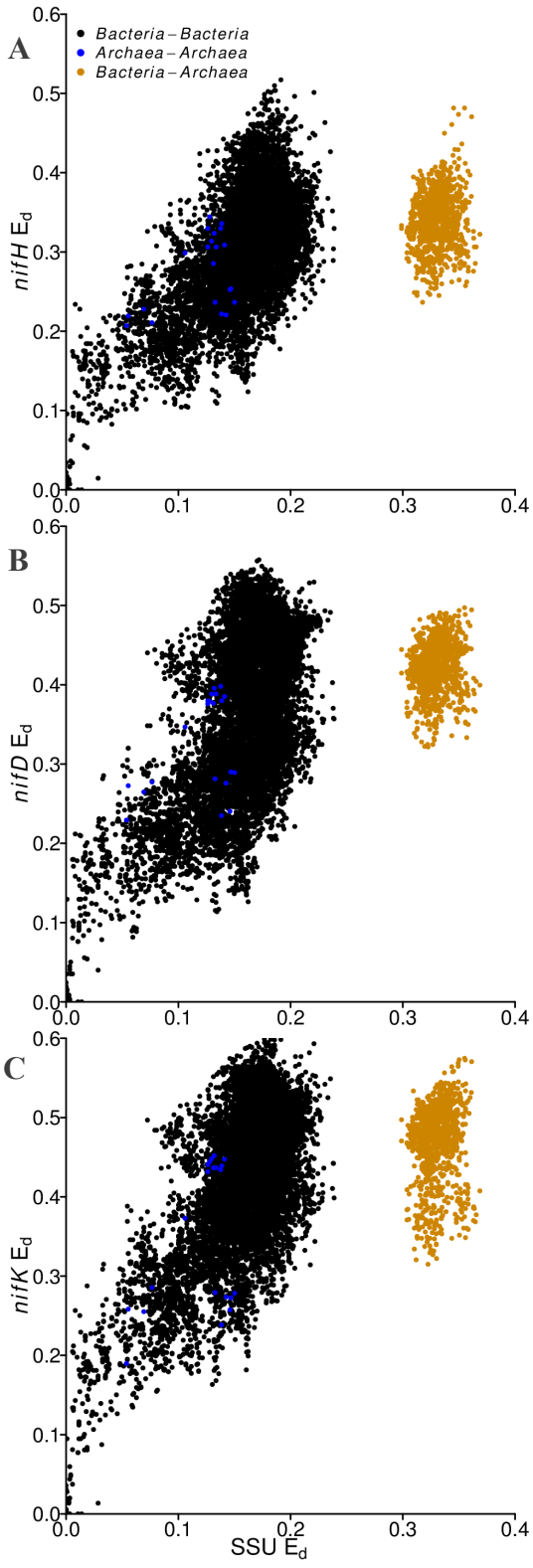
rate variation in *nifD* between cluster I and cluster III diazotrophs (Figure 2.8). Cluster I *nifD* sequences accumulate mutations slower than cluster III *nifD* sequences when compared to either *nifK* or *nifH* and these results are significant as determined by testing for homogeneity of slopes (*nifK*:f = 328.4, $p < 2 \times 10^{-16}$; *nifH*: f = 294.3, $p < 2 \times 10^{-16}$; Figure 2.8). In contrast, there is no difference between cluster I and cluster III diazotrophs with respect to the relative rate at which mutations accumulate in *nifD* and *nifK* (f = 3.451, p = 0.0633; Figure 2.8C). These data indicate that Cluster I diazotrophs accumulate mutations in *nifD* at a rate that is slower than that observed for cluster III diazotrophs.

Lastly, it is generally assumed that the *nifH* gene is the more highly conserved of the three structural *nif* genes and this is the basis of its use as a marker gene. To test this assumption, we evaluated *nif* genes for the 185 genomes. The sequence dissimilarity from the distance matrices of the pairwise sequence comparisons was 0.306 ± 0.069 for *nifH*, 0.363 ± 0.097 for *nifD*, and 0.421 ± 0.087 for *nifK* (mean ± standard deviation). Two-sample t-tests between *nifH* and *nifD*, and between *nifH* and *nifK* were both highly significant with $p<2.2 \times 10^{-16}$. This result confirms that *nifH* is on average more conserved than the other *nif* genes.

Comparative analysis of *nif* and 16S rRNA genes

Evolutionary distances (Ed) were calculated by pairwise comparison of 16S

Figure 2.9 Comparison of evolutionary distances (Ed) between16S rRNA genes and *nifH* (A), *nifD* (B), and *nifK* (C). Distances are calculated after exclusion of both indel regions and regions of dubious positional homology Colors indicate intra-domain and inter-domain comparisons for each set of genes as indicated in the figure legend. The number of sequences compared and details of distance calculation are as described in Figure 2.8.


A
B
C
Bacteria – Bacteria
Archaea – Archaea
Bacteria – Archaea
nifH Ed
nifD Ed
nifK Ed
SSU Ed
0.0
0.1
0.2
0.3
0.4
0.5
0.6

rRNA and *nif* genes from the 185 genomes (Figure 2.9). There is clear evidence for horizontal gene transfer of *nif* genes between *Bacteria* and *Archaea* (Figure 2.9). *Archaea-Bacteria* comparisons clearly partition out from intra-domain comparisons, and this can only occur as the result of transfer of *nif* genes between *Bacteria* and *Archaea*, though the plots do not allow inference of the direction of transfer (Figure 2.9).

We also evaluated *nifH* sequence divergence among genomes that share greater than 97% 16S rRNA sequence similarity. Microbial species are often delineated using a 97% 16S rRNA gene sequence similarity threshold [52]. Genomes that have 16S rRNA genes that are more than 97% similar can have up to 20% dissimilarity in their *nifH* sequences (Figure 2.10).

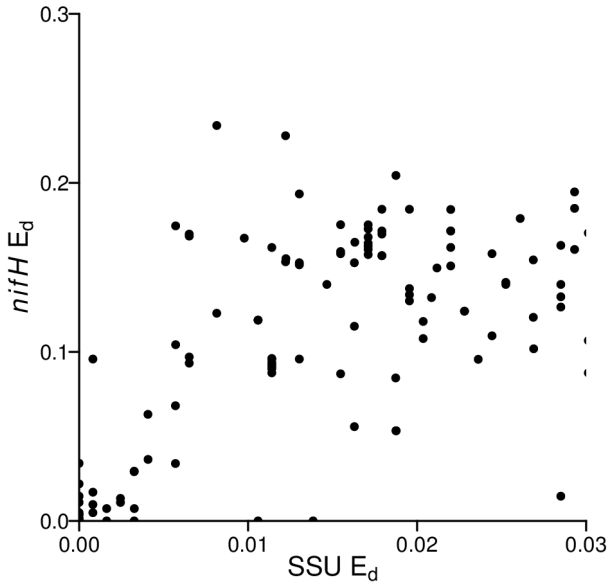


Figure 2.10 Comparison of evolutionary distances (Ed) between 16S rRNA and *nifH* genes from sequenced genomes indicating levels of *nifH* divergence found among 16S rRNA gene sequences that have less than 3% sequence difference.

Discussion

nifH sequencing trends

Patterns of *nifH* sequence acquisition over time provide insight on the composition of the current *nifH* database. The first *nifH* nucleotide sequence was acquired in 1980 from the cyanobacterium *Anabaena* 7120 [53], and the first survey of *nifH* sequences from the environment came in 1989 [54]. Since then a wide range of environments have been explored for nitrogenase diversity, and the increase in the number of sequences acquired per year (Figure 2.1) follows the trend for sequencing in general [55]. The *nifH* gene was the fourth most widely sequenced microbial functional gene as of July 12, 2012. While the majority of *nifH* sequence fragments contributed to the Genbank nucleotide database are obtained by Sanger-sequencing, the emergence of next generation sequencing technologies will likely yield a several-fold increase in the rate of sequence acquisition. Notably, the most common length for *nifH* amplicons is 324 nucleotides (Figure 2.4) which is suitable for 454 pyrosequencing, and a few studies have already taken advantage of *nifH* pyrosequencing [9, 56]. Inclusion in the database of sequences obtained through 454 technology would drastically increase its size, but it could also increase the error content of the reads relative to Sanger sequencing [57]. The *nifH* database will benefit in the future from the availability of more full-length sequences as more genome

sequences are determined for nitrogen-fixing organisms.

Sampling evenness

Although the *nifH* sequences in the database have been drawn from a wide range of environments, it is clear that sequences from terrestrial habitats and the open ocean are overrepresented among determined *nifH* sequences (Table 2.1). Furthermore, about half of the *nifH* sequences from terrestrial sites are attributable to agricultural soils. Nitrogenase *nifH* sequences from extreme environments, insect and mammal guts, compost, and anthropogenic environments such as water treatment and distribution systems are poorly represented in the database (Table 2.1). In addition, we observe that relatively few studies have focused on anoxic environments. Despite the few studies performed on anoxic environments, a relatively large number of sequences can be found in sequence cluster III which contains exclusively anaerobic organisms. In addition, this anaerobe-containing *nifH* sequence cluster contains very high levels of sequence diversity [45]. The process of nitrogen fixation may have evolved early in the history of life, when the Earth was anoxic [58]. The nitrogenase enzyme is oxygen sensitive, and thus anoxic environments have great potential for harboring *nifH* sequence diversity.

The limited geographic and habitat coverage in the database has likely prevented recovery of the full phylogenetic diversity of nitrogen fixers. However, it is

important to keep in mind that prior sampling has been governed by site specific research goals and has not been informed by global survey information. In addition, within the database there is great discrepancy in the number of sequences contributed from each study. For example, the top 1% of studies account for more than a quarter of all sequence depositions (Figure 2.2), and all originate after 2005. It will be valuable to obtain samples of *nifH* genes from geographically and ecologically disparate locations in order to obtain a complete understanding of nitrogenase evolution and diversity.

Annotation lacking

Many *nifH* sequences in public databases lack adequate metadata. For instance, 4,521 or 13.7% of entries lack data for "isolation source". In addition, only 6,796 or 20.6% of entries were annotated for latitude-longitude coordinates. Furthermore, records lack a basic description of useful environmental variables such as pH, moisture, organic matter, salinity, temperature, depth, etc. Meager annotation is a recognized problem [60, 68] because it limits the value of sequence information beyond the study of origin. Fortunately, efforts to standardize the recording of sequence metadata have progressed [59].

Comparison with other *nifH* gene compilations

There are two other collections of aligned *nifH* sequences which are currently publicly available. The Functional Gene Pipeline/Repository, known as FunGene, [61] http://fungene.cme.msu.edu//index.spr) provides an aligned collection of *nifH* sequences. The sequences in FunGene are pulled from public databases and aligned using HMMer. The FunGene sequence compilations are not in ARB database format and this data is not incorporated into a phylogenetic context and lacks sequence-associated metadata. A second *nifH* sequence collection is available from the Zehr research group (http://pmc.ucsc.edu/~wwwzehr/research/database/) in ARB database format. This database is compiled using a set of representative NifH protein sequences with BLAST to identify other NifH in the public repositories. A Hidden Markov Model approach via the software HMMer [62]; http://hmmer.janelia.org/) is used to align the sequences in the Zehr database. Putative chimeric sequences within the Zehr database were identified with UCHIME [63] and labeled, and the program CD-HIT [64] was used to establish sequence clusters. Unlike the Zehr database, our database includes linked *nifH, nifD, nifK,* and 16S rRNA gene records from 185 sequenced genomes, which allows for exploration of evolutionary relationships (Figures 2.8 and 2.9). Our database also includes a comprehensive tree of all *nifH* sequences and trees of the full-length sequences for *nifH*, *nifD*, *nifK*, and 16S rRNA genes. These trees can be used to organize and navigate sequence data.

Using the database to evaluate nitrogenase evolution

The evolutionary history of nitrogen fixation has been the subject of numerous studies which have explored the emergence of *nif* genes and related paralogs. Early studies noted that *nifD* and *nifK* are homologous [65], leading to the inference supported by phylogenetic analysis that *nifK* is the result of a duplication of *nifD*. Additional analysis showed that *nifEN* was the result of an in tandem duplication of *nifDK* [66, 67]. The alternative nitrogenases encoded by the *vnf* cluster, which encodes the V-Fe nitrogenase, and the *anf* cluster, which encodes the Fe-Fe nitrogenase, were initially thought to have emerged prior to the Mo-Fe nitrogenase based upon the knowledge that the early Earth was anoxic and that the availability of Mo would have been limited in an anoxic ocean [68]. However, based upon phylogenetic analysis of *nif, vnf, anf,* and another set of paralogous genes encoding bacteriochlorophyll synthesis, the Mo-Fe nitrogenase was shown to have emerged as the first nitrogenase [69], followed by *vnf*, then *anf*. NifEN acts as a scaffold for synthesis of the Mo-Fe cofactor which allows for more efficient nitrogen fixation, and the emergence of NifEN followed relatively shortly after the emergence of oxygen in the Earth's atmosphere [69]. With regard to explanations for the present day phylogenetic occurrence of nitrogen fixation, two possibilities were proposed. One is that the last common ancestor was capable of nitrogen fixation and that there has been loss of function among many taxa, while the other possibility invokes horizontal transfer. The

later explanation has been supported by phylogenetic analysis which affirms that nitrogen fixation first emerged in the *Methanococcales* and *Methanobacteriales* and was later horizontally transfered to an ancestor of the *Firmicutes* from an ancestor of the *Methanosarcinales* [69].

We used the genome entries in our database to compare the evolutionary distances of *nif* genes relative to 16S rRNA genes. This approach can demonstrate cases where horizontal gene transfer has occurred [70], evident as discontinuity on a plot of Ed values between a control and experimental gene (e.g. Figure 2.9). We observe partitioning between intra-domain and inter-domain *(Archaea* to *Bacteria)* comparisons of *nifH*, *nifD*, *nifK* relative to 16S rRNA genes (Figure 2.9). Such partitioning indicates horizontal gene transfer of the nitrogen-fixation genes from *Bacteria* to *Archaea* [70]. This observation is consistent with previous evidence of such a transfer event [71, 72] which has been recently confirmed [69].

We also compared Ed among *nif* genes to examine incongruencies among the two principal *nif* clusters, cluster I and cluster III. There is partitioning of Ed values between *nif* genes from cluster I and cluster III (Figure 2.8). The Ed comparison for *nifD* and *nifK* (Figure 2.8 C) shows a consistent accumulation of mutations over evolutionary time. Ed comparisons that include *nifH* (Figure 2.8 A and B) indicate discontinuity in the accumulation of mutations between cluster I and cluster III. This discontinuity indicates that *nifH* has a distinct evolutionary history from *nifD* and *nifK*. Furthermore, this pattern of partitioning is consistent with *nifH* having been

transferred between cluster I and III at a point after the emergence of *nifDK*. That *nifD* and *nifK* show distinct evolutionary histories from *nifH* is consistent with the fact that *nifD* and *nifK* are constrained to a high degree by their tight structural association in the core of the nitrogenase enzyme. In contrast, *nifH* interacts with a limited number of residues on the exterior of the core nitrogenase enzyme. Previous research has shown that *nifD* and *nifK* are ancient paralogs [65], and thus have a common evolutionary history. We show that as a result of horizontal gene transfer, *nifH* has evolutionary origins that are different from those of *nifDK*.

Analysis of Ed values also indicated that the rate of mutational accumulation in *nifD* differs between cluster I and cluster III sequences (Figure 2.8 A and C). In contrast, the rate at which mutations accumulated in *nifH* and *nifK* did not differ between cluster I and cluster III sequences. Differences in the rate of mutation fixation between gene sequence clusters indicates a difference in the evolutionary forces that shape protein evolution and are expected for proteins under different functional constraints. Models of nitrogenase protein structure indicate that the cluster III type nitrogenase interacts with a FeMo co-factor as would be expected if these genes are orthologous with cluster I type FeMo nitrogenase (*nif*) rather than either of the FeFe (*anf*) or FeV (*vnf*) alternative nitrogenases [73]. In *nifD*, the active site where cofactor binding occurs displays sequence conservation according to the type of metal cofactor bound, with cluster I and III showing similar residues though cluster III harbors an internal amino acid extension which distinguishes it structurally from

cluster I and the alternative nitrogenases [73]. Another major distinguishing feature of cluster III versus cluster I is that cluster III contains strict anaerobes exclusively while cluster I contains both aerobic organisms and facultative anaerobes. In addition, cluster III FeMo nitrogenases are likely ancestral to both cluster I FeMo nitrogenase and the alternative FeFe and VFe nitrogenases [69]. That mutations accumulate at different rates in the cluster III and cluster I type nitrogenases suggests that NifD is under different evolutionary constraints in these two groups. Further work is needed to determine the functional significance of NifD residues that differ between the cluster I and cluster III type nitrogenases and to determine whether these enzymes are in fact functionally equivalent.

Lastly, our examination of evolutionary distance in *nifH* has consequences for diversity analyses that use the *nifH* gene as a marker gene. Sequence-based assessments of microbial diversity often use a similarity cutoff to delineate operational taxonomic units (OTU) that are thought to correspond to species. A 97% similarity cutoff is often applied to the 16S rRNA gene to define species level OTUs [74]. Genome sequence analysis indicates that species level OTUs can be defined using a 95% similarity cutoff for conserved protein encoding genes [75]. However, we find that 16S rRNA gene sequences that share 97% similarity have *nifH* evolutionary distance values that range from 0 to 0.2 (Figure 2.10). Hence, two strains that belong to the same OTU as defined by their 16S rRNA genes can have *nifH* genes that differ in 20% of nucleotide positions. Hence, measures of diversity based on *nifH* are not

likely to be directly comparable to measures of diversity made with 16S rRNA genes. Furthermore, while it can be informative to discuss sequence diversity in *nifH* it is not possible to use *nifH* sequence diversity as an estimate of species diversity.

Summary

The database we describe includes all *nifH* sequences present in the Genbank nucleotide database as of May 16, 2012 and includes associated metadata in a searchable framework. The database is available in ARB format, allowing for point-and-click access to sequence record importation and exportation, searching, alignment editing, primer design and matching, and phylogenetic tree construction [38]. The *nifH* sequences are organized into phylogenetic trees that can be navigated and used for sequence selection. Users may find the database useful for diversity analyses [95], primer evaluations [28], examination of potential sampling and habitat bias, G+C content analysis (Figure 2.6 and 2.7), characterization of horizontal gene transfer events (e.g. [76], Figure 2.9), and phylogeny analysis.

References

1. Vitousek P, Aber J, Howarth R, Likens G, Matson P et al. (1997) Human alteration of the global nitrogen cycle: sources and consequences. Ecol Appl 7: 750.

2. Peoples MB, Craswell ET (1992) biological nitrogen-fixation: investments, expectations and actual contributions to agriculture. Plant Soil 141: 13-39.

3. Kennedy I, Islam N (2001) The current and potential contribution of asymbiotic nitrogen fixation to nitrogen requirements on farms: a review. Aust J Exp Ag 41: 457.

4. Cleveland C, Townsend A, Schimel D, Fisher H, Howarth R et al. (1999) Global patterns of terrestrial biological nitrogen (N-2) fixation in natural ecosystems. Global Biogeochemical Cycles 13: 645.

5. Benton, PMC, Sen, S, and Peters, JW (2002) Nitrogenase Structure. In: Leigh JG, editor. Nitrogen fixation at the millennium. Elsevier Science B.V.. pp. 35-71.

6. Young JPW (2005) The phylogeny and evolution of nitrogenases. In: Palacios R, Newton WE, editors. Genomes and genomics of nitrogen-fixing organisms. Springer. pp. 221-241.

7. Fisher K, Newton WE (2002) Nitrogen fixation--a general overview. In: Leigh GJ, editor. Nitrogen fixation at the millennium. Elsevier Science B.V.. pp. 1-34.

8. Raymond J, Siefert JL, Staples CR, Blankenship RE (2004) The natural history of nitrogen fixation. Mol Biol Evol 21: 541-554.

9. Farnelid H, Andersson AF, Bertilsson S, Al-Soud WA, Hansen LH et al. (2011) Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. PLoS One 6: e19223.

10. Langlois RJ, LaRoche J, Raab PA (2005) Diazotrophic diversity and distribution in the tropical and subtropical Atlantic ocean. Appl Environ Microbiol 71: 7910-7919.

11. Zehr JP, Mellon MT, Zani S (1998) New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of Nitrogenase (*nifH*) genes. Appl Environ Microbiol 64: 3444-3450.

12. Roesch L, Camargo F, Bento F, Triplett E (2008) Biodiversity of diazotrophic bacteria within the soil, root and stem of field-grown maize. Plant Soil 302: 91-104.

13. Izquierdo JA, Nüsslein K (2006) Distribution of extensive *nifH* gene diversity across physical soil microenvironments. Microb Ecol 51: 441-452.

14. Rösch C, Mergel A, Bothe H (2002) Biodiversity of denitrifying and dinitrogen-fixing bacteria in an acid forest soil. Appl Environ Microbiol 68: 3818-3829.

15. Mehta MP, Butterfield DA, Baross JA (2003) Phylogenetic diversity of nitrogenase (*nifH*) genes in deep-sea and hydrothermal vent environments of the Juan de Fuca Ridge. Appl Environ Microbiol 69: 960-970.

16. Zehr JP, Jenkins BD, Short SM, Steward GF (2003) Nitrogenase gene diversity and microbial community structure: a cross-system comparison. Environ Microbiol 5: 539-554.

17. Gaby JC, Buckley DH (2011) A global census of nitrogenase diversity. Environ Microbiol 13: 1790-1799.

18. Hsu S, Buckley DH (2009) Evidence for the functional significance of diazotroph community structure in soil. ISME J 3: 124-136.

19. Chien Y, Zinder SH (1994) Cloning, DNA-sequencing, and characterization of a *nifD*-homologous gene from the archaeon *Methanosarcina barkeri* 227 which resembles *nifD1* from the eubacterium *Clostridium pasteurianum*. J Bacteriol 176: 6590-6598.

20. Masepohl B, Schneider K, Drepper T, Müller A, Klipp W (2002) Alternative nitrogenases. In: Leigh GJ, editor. Nitrogen fixation at the millennium. Elsevier Science B.V. pp. 191-222.

21. Souillard N, Magot M, Possot O, Sibold L (1988) Nucleotide sequence of regions homologous to *nifH* (nitrogenase Fe protein) from the nitrogen-fixing archaebacteria *Methanococcus thermolithotrophicus* and *Methanobacterium ivanovii*: evolutionary implications. J Mol Evol 27: 65-76.

22. Staples CR, Lahiri S, Raymond J, Von Herbulis L, Mukhophadhyay B et al. (2007) Expression and association of group IV nitrogenase NifD and NifH homologs in the non-nitrogen-fixing archaeon *Methanocaldococcus jannaschii*. J Bacteriol 189: 7392-7398.

23. Fujita Y, Takahashi Y, Chuganji M, Matsubara H (1992) The *nifh*-like (*frxc*) gene is involved in the biosynthesis of chlorophyll in the filamentous cyanobacterium *Plectonema boryanum*. Plant Cell Physiol 33: 81-92.

24. Nomata J, Mizoguchi T, Tamiaki H, Fujita Y (2006) A second nitrogenase-like enzyme for bacteriochlorophyll biosynthesis - Reconstitution of chlorophyllide a reductase with purified X-protein (BchX) and YZ-protein (BchY-BchZ) from *Rhodobacter capsulatus*. J Biol Chem 281: 15021-15028.

25. Schloss P, Handelsman J (2004) Status of the microbial census. Microbiol Mol Biol Rev 68: 686-691.

26. Rappe M, Giovannoni S (2003) The uncultured microbial majority. Annu Rev Microbiol 57: 369-394.

27. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol 33(2): reviews0003.1–0003.8.

28. Gaby JC, Buckley DH (2012) A Comprehensive Evaluation of PCR Primers to Amplify the *nifH* Gene of Nitrogenase. PLoS One 7: e42149.

29. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA et al. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res 33: D294-6.

30. Moisander PH, Beinart RA, Hewson I, White AE, Johnson KS et al. (2010) Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. Science 327: 1512-1514.

31. Wallenstein MD, Vilgalys RJ (2005) Quantitative analyses of nitrogen cycling genes in soils. Pedobiologia 49: 665-672.

32. Pereira e Silva MC, Semenov AV, van Elsas JD, Salles JF (2011) Seasonal variations in the diversity and abundance of diazotrophic communities across soils. FEMS Microbiol Ecol 77: 57-68.

33. Martensson L, Diez B, Wartiainen I, Zheng W, El-Shehawy R et al. (2009) Diazotrophic diversity, *nifH* gene expression and nitrogenase activity in a rice paddy field in Fujian, China. Plant and Soil 325: 207-218.

34. Giebel H, Kalhoefer D, Lemke A, Thole S, Gahl-Janssen R et al. (2011) Distribution of Roseobacter RCA and SAR11 lineages in the North Sea and characteristics of an abundant RCA isolate. ISME J 5: 8-19.

35. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188–7196.

36. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72: 5069-5072.

37. Cole JR, Wang Q, Cardenas E, Fish J, Chai B et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37: D141-5.

38. Ludwig W, Strunk O, Westram R, Richter L, Meier H et al. (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32: 1363-1371.

39. Finn RD, Mistry J, Tate J, Coggill P, Heger A et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211-22.

40. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41: 95-98.

41. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. (2007) ClustalW and ClustalX version 2.0. Bioinformatics, 23: 2947-2948.

42. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72: 5069-5072.

43. United Nations. Standard Country or Area Codes for Statistical Use. Retrieved February 23, 2013, from http://unstats.un.org/unsd/methods/m49/m49.htm.

44. United Nations (1999) Standard country or area codes for statistical use (current information as at 31 august 1999). United Nations, New York.

45. Team RC (2012) R: A Language and Environment for Statistical Computing.

46. Team RC (2012) R: A Language and Environment for Statistical Computing.

47. Severin I, Stal LJ (2010) NifH expression by five groups of phototrophs compared with nitrogenase activity in coastal microbial mats. FEMS Microbiol Ecol 73: 55-67.

48. Fouts DE, Tyler HL, DeBoy RT, Daugherty S, Ren Q et al. (2008) Complete genome sequence of the N_2-fixing broad host range endophyte *Klebsiella pneumoniae* 342 and virulence predictions verified in mice. PLoS Genet 4: e1000141.

49. Jacobson MR, Brigle KE, Bennett LT, Setterquist RA, Wilson MS et al. (1989) Physical and genetic map of the major nif gene cluster from *Azotobacter vinelandii*. J Bacteriol 171: 1017-1027.

50. Arnold W, Rump A, Klipp W, Priefer UB, Pühler A (1988) Nucleotide sequence of a 24,206-base-pair DNA fragment carrying the entire nitrogen fixation gene cluster of *Klebsiella pneumoniae*. J Mol Biol 203: 715-738.

51. Farahi K, Pusch GD, Overbeek R, Whitman WB (2004) Detection of lateral gene transfer events in the prokaryotic tRNA synthetases by the ratios of evolutionary distances method. J Mol Evol 58: 615-631.

52. Stackebrandt, E. and Goebel, B.M. (1994) A place for DNA-DNA reassociation and 16s ribosomal-RNA sequence-analysis in the present species definition in bacteriology. Int J Syst Bacteriol 44: 846-849.

53. Mevarech M, Rice D, Haselkorn R (1980) Nucleotide sequence of a cyanobacterial *nifH* gene coding for nitrogenase reductase. Proc Natl Acad Sci U S A 77: 6476-6480.

54. Zehr JP, McReynolds LA (1989) Use of degenerate oligonucleotides for amplification of the *nifH* gene from the marine cyanobacterium *Trichodesmium thiebautii*. Appl Environ Microbiol 55: 2522-2526.

55. Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour 8: 3-17.

56. Mao Y, Yannarell AC, Mackie RI (2011) Changes in N-transforming *Archaea* and *Bacteria* in soil during the establishment of bioenergy crops. PLoS One 6: e24750.

57. Shendure J, Mitra RD, Varma C, Church GM (2004) Advanced sequencing technologies: methods and goals. Nat Rev Genet 5: 335-344.

58. Navarro-González R, McKay CP, Mvondo DN (2001) A possible nitrogen crisis for Archaean life due to reduced nitrogen fixation by lightning. Nature 412: 61-64.

59. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR et al. (2011) Minimum

information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 29: 415-420.

60. (2008) A place for everything. Nature 453: 2.

61. Chai B, Kulam S, McGarrell DM, Wang Q, Farris RJ et al. (2005) The functional gene pipeline/repository. Abstracts of the General Meeting of the American Society for Microbiology 105.

62. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7: e1002195.

63. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27: 2194-2200.

64. Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26: 680-682.

65. Thony B, Kaluza K, Hennecke H (1985) structural and functional homology between the alpha-subunits and beta-subunits of the nitrogenase MoFe protein as revealed by sequencing the *Rhizobium japonicum nifK* gene. Mol Gen Genet 198: 441-448.

66. Dean DR, Brigle KE (1985) *Azotobacter vinelandii nifD*- and *nifE*-encoded polypeptides share structural homology. Proc Natl Acad Sci U S A 82: 5720-5723.

67. Brigle KE, Weiss MC, Newton WE, Dean DR (1987) Products of the iron-molybdenum cofactor-specific biosynthetic genes, *nifE* and *nifN,* are structurally homologous to the products of the nitrogenase molybdenum-iron protein genes, *nifD* and *nifK*. J Bacteriol 169: 1547-1553.

68. Scott C, Lyons TW, Bekker A, Shen Y, Poulton SW, Chu X et al. (2008) Tracing the stepwise oxygenation of the Proterozoic ocean. Nature 452: 456-459.

69. Boyd ES, Anbar AD, Miller S, Hamilton TL, Lavin M, Peters JW (2011) A late methanogen origin for molybdenum-dependent nitrogenase. Geobiology 9: 221-232.

70. Farahi K, Pusch GD, Overbeek R, Whitman WB (2004) Detection of lateral gene transfer events in the prokaryotic tRNA synthetases by the ratios of

evolutionary distances method. J Mol Evol 58: 615-31.

71. Chien Y, Zinder SH (1994) Cloning, DNA-sequencing, and characterization of a *nifD-homologous* gene from the archaeon *Methanosarcina barkeri* 227 which resembles *nifD1* from the eubacterium *Clostridium pasteurianum*. J Bacteriol 176: 6590-6598.

72. Raymond J, Siefert JL, Staples CR, Blankenship RE (2004) The natural history of nitrogen fixation. Mol Biol Evol 21: 541-554.

73. Mcglynn, S.E., Boyd, E.S., Peters, J.W. and Orphan, V.J. (2013) Classifying the metal dependence of uncharacterized nitrogenases. Frontiers in Microbiology 3: doi: 10.3389/fmicb.2012.00419.

74. Hagstrom A, Pinhassi J, Zweifel UL (2000) Biogeographical diversity among marine bacterioplankton. Aquat Microb Ecol 21: 231-244.

75. Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci 361: 1929-40.

76. Farahi K, Whitman WB, Kraemer ET (2003) RED-T: utilizing the Ratios of Evolutionary Distances for determination of alternative phylogenetic events. Bioinformatics 19: 2152-2154.

CHAPTER 3

A GLOBAL CENSUS OF NITROGENASE DIVERSITY

Introduction

Nitrogen limits primary production in both terrestrial and marine ecosystems [1], and biological nitrogen fixation is the dominant natural process by which ecosystems obtain nitrogen. Nitrogen fixation is mediated solely by *Bacteria* and *Archaea* that possess the enzyme nitrogenase. The nitrogenase enzyme complex is encoded by the genes *nifH*, *nifD* and *nifK*, with *nifH* encoding the dinitrogenase reductase subunit (as reviewed in [2]). Zehr and McReynolds were the first to develop PCR primers for amplification of *nifH* genes, which they used to examine *Trichodesmium thiebautii* distribution and diversity in the Caribbean Sea [3]. This approach was later applied using diverse *nifH* primer sets to target a wide range of environments (as reviewed in [4]). The *nifH* gene has remained the signature gene for studying the diversity of nitrogen-fixing organisms. Zehr and colleagues [4] performed the last major evaluation of *nifH* sequences in public databases, using the approximately 1500 *nifH* sequences then available. In that effort the authors examined the phylogeny of *nifH* sequences and performed a cross-system comparison of the distribution of *nifH* lineages across different environments [4]. We have used the large and ever-growing body of *nifH* sequences available in public databases to determine the degree to which current sampling efforts have captured the global diversity of

nitrogen-fixing organisms, and to assess our current understanding of diazotroph diversity in different environments and in different *nifH* lineages.

Five major clusters with homology to *nifH* have been described [4-6, 10]. The majority of known *nifH* sequences fall into cluster I. Cluster I is composed entirely of *nifH* genes from the conventional FeMo nitrogenase of bacteria. This cluster contains genes from most *Proteobacteria,* all *Cyanobacteria* and certain *Firmicutes* (*Paenibacillus*) and *Actinobacteria* (*Frankia*) [4]. Cluster II contains a relatively small number of sequences that belong to the alternative FeV and FeFe nitrogenases as well as sequences belonging to certain methanogenic *Archaea* [4]. Cluster III is dominated by *nifH* sequences from anaerobic members of the *Bacteria* and *Archaea* including: spirochetes, methanogens, acetogens, sulfate-reducing bacteria, green sulfur bacteria and clostridia [4]. Clusters IV and V are composed of *nifH* paralogues that are not involved in nitrogen fixation and include genes of various functions including some involved in photopigment biosynthesis and certain electron transport reactions [8]. The phylogenies of *nifD, K*, *E* and *N* all generally agree with the nitrogenase clusters that have been determined using *nifH* [8].

Any effort to make cross-system comparisons of diversity are confronted by a variety of potential biases, and it is important to consider these constraints before interpreting analyses made at this scale. First, PCR primer bias can impact the diversity and relative abundance of *nifH* genes detected. Such biases can either exclude discovery of certain lineages [16] or can alter the ratios of sequence abundance in the products relative to the templates of PCR [10, 11]. Primer sets for *nifH* are designed to be either universal [12-14] or group-specific [9], but in practice each primer set exhibits a unique range of specificities. Because each primer set has its

own specificity or bias, and different research groups may favour different sets of primers and different PCR reaction conditions, there exists considerable potential for variation in results between laboratories. A second concern is that sampling efforts have been uneven with many studies performed on soils and the photic zone of the ocean and fewer performed on extreme or anoxic environments (sediments, microbial mats, gut contents, etc.). We have chosen to focus our analyses on taxonomic richness and have chosen to use the Chao1 richness estimator for making diversity estimates. The value of the Chao1 estimate is calculated using the frequency of sequences that occur exactly one or two times. This estimator has the disadvantage of providing richness estimates that are strongly dependent on the number of sequences analysed, but offers the advantage of providing confidence intervals that allow robust inter-sample comparisons provided that the same number of sequences is sub-sampled from each collection of sequences [15].

The existence of *nifH* paralogues that can be misannotated as *nifH* is also a cause for concern when conducting analyses of putative *nifH* sequences obtained from environmental sources. These paralogues can be split into two groups: those encoding subunits of alternative nitrogenases, and those not involved in nitrogen fixation. While the conventional nitrogenase contains a FeMo metal cluster, the alternative nitrogenases contain either FeV or FeFe metal clusters and the dinitrogenase reductase subunits of these enzymes are encoded, respectively, by *vnfH* and *anfH.* While *anfH* sequences fall into cluster II, *vnfH* sequences do not form a distinct phylogenetic cluster apart from *nifH* [24], and in the current study no effort was made to discriminate the *vnfH* sequences from the *nifH* sequences. Additionally, a range of *nifH* paralogues are commonly mis-annotated as *nifH* in sequence databases and

assembled genomes although they have no role in nitrogen fixation [16, 17]. When subject to phylogenetic analysis these paralogues fall out into cluster IV and V and can clearly be resolved from true nitrogenases [8]; these sequences have been excluded from our *nifH* diversity analyses.

In this report we analyse the current status of the *nifH* gene census through creation of a phylogenetically organized database of 16 989 *nifH* sequences. We used the database to make estimates of nitrogenase richness and estimates of coverage in different lineages and in different environments. This synthesis of sequence information provides context on our current understanding of nitrogenase diversity that can be used to direct future studies or formulate new hypotheses. Given the limitations and biases associated with examining data present in public databases it is important to recognize that this report cannot make conclusive statements about absolute differences in diversity. This report provides a glimpse at the current status of the census for nitrogen-fixing organisms and characterizes our current understanding of diazotroph diversity.

Results

Richness estimates for environmental categories

The majority of available *nifH* sequences provide only partial coverage of the gene and many of these partial sequences do not overlap completely. To address this issue we calculated diversity estimates using the range of columns in the sequence alignment that provided the greatest degree of overlap, consisting of 10 833 sequences

that spanned a region of 322 nucleotide positions (as described in Experimental procedures). These 10 833 sequences comprised 8193 unique sequences, 3358 $OTU_{0.05}$, 2341 $OTU_{0.10}$, 809 $OTU_{0.20}$ and 23 $OTU_{0.40}$ (Figure 3.1). The accumulation curves at $OTU_{0.20}$, and $OTU_{0.40}$ appeared to level off while those at other OTU cut-offs did not (Figure 3.1), indicating that we still have an incomplete census of diazotroph species but that sampling of the major lineages of diazotrophs is fairly complete. We observed that an $OTU_{0.40}$ roughly delineates the major *nifH* sequence clusters. The greatest number of these 10 833 sequences came from either soil (3644) or marine

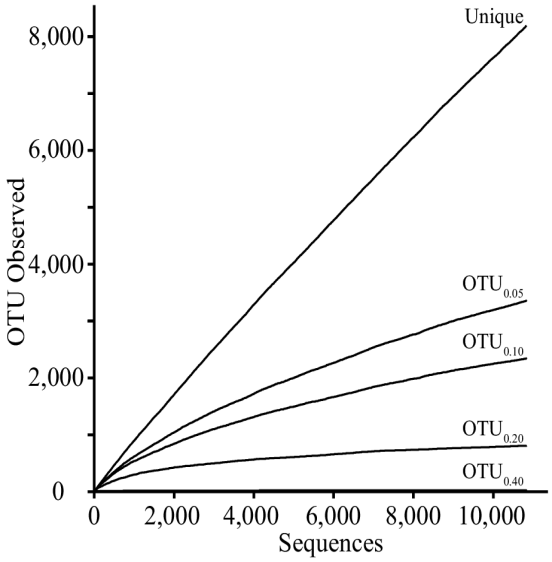


Figure 3.1: Collector's curves for 10 833 *nifH* sequences that shared a common frame of 322 nucleotides (*A. vinelandii* positions 133 to 454). Lines indicate the number of OTUs detected using different similarity cut-offs as indicated in the figure.

sources (2855). Soil contained a greater diversity of diazotrophs than any other environment represented in the database (Figure 3.2). When the analysis is limited to

2855 sequences to equalize sampling intensity for each sequence set, the Chao1 richness estimate for soil was 3216 $OTU_{0.05}$ [2864 lower 95% confidence interval (LCI), and 3646 higher 95% confidence interval (HCI)] and that for marine systems was 1884 $OTU_{0.05}$ (1581 LCI, 2286 HCI) and this difference is statistically significant. Likewise, the diversity of soil is still the highest when microbial mat communities are included in the analysis (which requires assessing richness at a sampling intensity of 1222 sequences) (Figure 3.2). At this sampling intensity the richness of microbial mats (890 $OTU_{0.05}$; 737 LCI, 1110 HCI) was not significantly different from that of marine

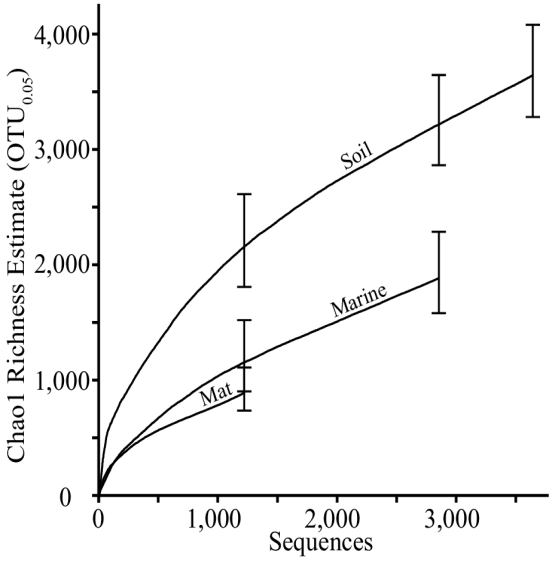


Figure 3.2: Chao1 richness estimates for *nifH* sequences belonging to different environmental categories. Bars indicate 95% confidence intervals plotted at the sampling endpoints of each curve. The environmental categories are indicated by labels in the figure.

systems (1154 OTU0.05; 903 LCI, 1522 HCI), but both of these estimates were lower than the richness estimate for soil (2156 $OTU_{0.05}$; 1809 LCI, 2612 HCI) and these differences are significant. The Chao1 richness estimator systematically

underestimates richness at low levels of sampling but permits robust comparisons of relative richness between environments [28], thus Chao1 richness estimates should be treated as a lower bound. Richness estimates for phylogenetic clusters Chao1 richness estimates were calculated for phylogenetic groups and the greatest sequence richness was observed in cluster III and the lowest for the *Cyanobacteria* (Figure 3.3). When assessed at common sampling intensity (2089 sequences), the Chao1 richness estimate for the *Cyanobacteria* (480 $OTU_{0.05}$; 410 LCI, 590 HCI) was significantly lower than that for the *α*, *β*, and *γ Proteobacteria* (1818 $OTU_{0.05}$; 1587 LCI, 2117 HCI), which in

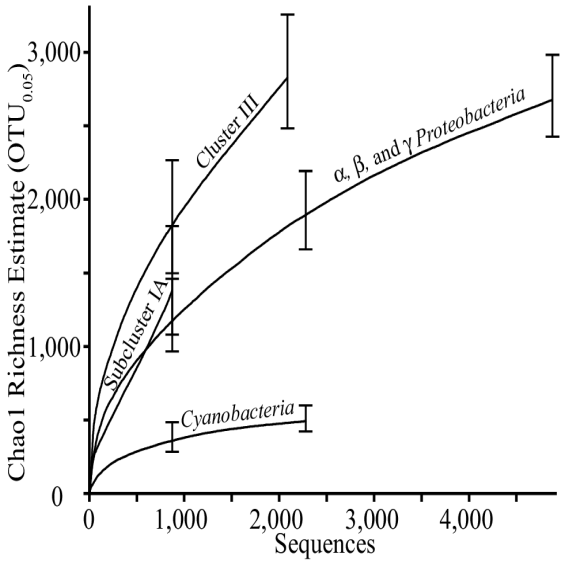


Figure 3.3: Chao1 richness estimates at $OTU_{0.05}$ for *nifH* sequences belonging to different phylogenetic clusters as indicated by labels in the figure. Selection of clusters was informed by the OTU groupings established by DOTUR analysis at $OTU_{0.40}$. Bars indicate 95% confidence intervals plotted at the sampling endpoints of each curve.

turn was significantly lower than that for cluster III (2829 $OTU_{0.05}$; 2483 LCI, 3259 HCI). Subcluster IA, which contains *δ Proteobacteria* of the *Desulfuromonadales*, is a distinct and divergent group that represents approximately 8% of the *nifH* sequences

in the database (Table 3.1). Thus, we treated subcluster IA as a distinct group in our analysis. To make comparisons of Chao1 richness estimates that include subcluster IA requires limiting the sampling intensity to 873 sequences. At this sampling intensity the Chao1 richness estimate for subcluster IA (1382 $OTU_{0.05}$; 1081 LCI, 1819 HCI) was significantly higher than that for the *Cyanobacteria* (358 $OTU_{0.05}$; 283 LCI, 485 HCI) but was not significantly different from estimates obtained for the α, β, and γ *Proteobacteria* or cluster III.

Richness was also evaluated at a deeper level of phylogenetic resolution ($OTU_{0.20}$). This cut-off provides an objective estimate for the number of major subgroups present within each of the larger clusters and subclusters examined.

Table 3.1: Count of *nifH* sequences as a function of phylogenetic cluster and source.

	Termite		Mat		Marine		Soil		Other[c]		Total[d]
	Seq.[a]	%[b]	Seq.[a]	%[b]	Seq.[a]	%[b]	Seq.[a]	%[b]	Seq.[a]	%[b]	Seq.[a]
Subcluster IA	0	0	17	1	143	4	822	14	364	6	1,346
Cluster III	300	57	571	39	395	13	644	11	798	13	2,708
Cyanobacteria	0	0	719	49	957	30	562	10	635	10	2,873
α, β, and γ *Proteobacteria*	6	1	148	10	1,437	46	3,219	56	3,203	53	8,013
Other[e]	223	42	12	1	209	7	501	9	1,104	18	2,049
Total[f]	529	100	1,467	100	3,141	100	5,748	100	6,104	100	16,989

[a] The number of *nifH* sequences from each *nifH* cluster in the environment specified.
[b] The percentage of *nifH* sequences from each *nifH* cluster in the environment specified.
[c] Indicates *nifH* data from environments not listed in preceding columns as well as sequences that could not be attributed to an environmental source.
[d] The total number of sequences in the database belonging to each *nifH* cluster.
[e] Indicates *nifH* data from sequence clusters not represented in preceding rows.
[f] Indicates the total number of sequences in the database associated with each environment.

When a uniform sampling intensity of 2089 sequences was applied, cluster III was estimated to contain by far the greatest level of diversity (639 $OTU_{0.20}$; 596 LCI, 702 HCI) followed by the *α*, *β*, and *γ Proteobacteria* (154 $OTU_{0.20}$; 138 LCI, 195 HCI) and *Cyanobacteria* (47 $OTU_{0.20}$; 44 LCI, 66 HCI) and these differences are significant. The richness of subcluster IA could only be assessed at a sampling intensity of 873 sequences (56 $OTU_{0.20}$; 40 LCI, 122 HCI). At this level of sampling the richness of subcluster IA did not differ significantly from that of the *Cyanobacteria,* but it was lower than both that of cluster III (566 $OTU_{0.20}$; 492 LCI, 676 HCI) and the *α*, *β*, and *γ Proteobacteria* (129 $OTU_{0.20}$; 114 LCI, 166 HCI) and these differences were significant. In addition, collector's curves generated using the $OTU_{0.20}$ criterion demonstrate that the discovery of new groups within the *α*, *β*, and *γ Proteobacteria*, the *Cyanobacteria,* and within subcluster IA is increasingly unlikely, while diversity within cluster III remains poorly sampled.

Rank-abundance distribution of *nifH* sequences

A rank-abundance plot made at $OTU_{0.05}$ for the 10 833 *nifH* sequences shows a long tail of rare sequences (Figure 3.4). When clustered at $OTU_{0.05}$, there are 2097 sequences that were observed only once, comprising 19% of the sequences analysed (Figure 3.4). The observation of a long tail of rare species was likewise observed for most clusters when frequency distributions were expressed for individual phylogenetic clusters (Figure 3.5). The OTU frequency distribution for the *Cyanobacteria,* however, displayed a striking pattern of dominance, with a small number of dominant taxa, which distinguished it from the other phylogenetic clusters (Figure 3.5). This

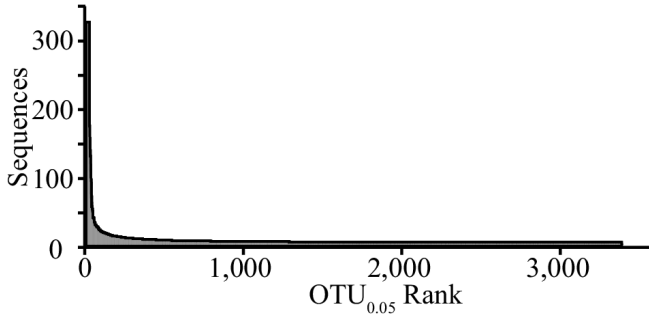


Figure 3.4: Rank-abundance distribution calculated for all $OTU_{0.05}$ in the *nifH* database.

pattern of dominance can be expressed as a reduction in evenness and quantified using Pielou's J. As expected, the value of Pielou's J for the *Cyanobacteria* (0.727) was lower than that of any other cluster [subcluster IA (0.922), *α*, *β*, and *γ Proteobacteria* (0.857), and cluster III (0.932)] consistent with the strong pattern of dominance.

We also identified the nitrogen-fixing taxa (defined at $OTU_{0.05}$) that are most abundant in the database (Tables 3.2). These data are strongly influenced by site selection bias in the surveys conducted to date, and may also be influenced by PCR-related biases, and they should not be taken as a measure of global abundance. Regardless, they provide information on the nitrogen-fixing organisms most commonly observed in sequence databases. Five of the 10 most observed taxa (at $OTU_{0.05}$) are from the *Cyanobacteria,* and the other five are from the *α*, *β*, and *γ Proteobacteria.* These 10 taxa represent a total of 1601 sequences, or 15% of the 10 833 sequences assessed with 864 sequences (8%) belonging to *Cyanobacteria* and 737 sequences (7%) belonging to *Proteobacteria.* Five of the 10 OTUs do not contain cultivated representatives, and six of the 10 were observed primarily in marine systems. The most frequently observed OTU (consisting of 320 sequences

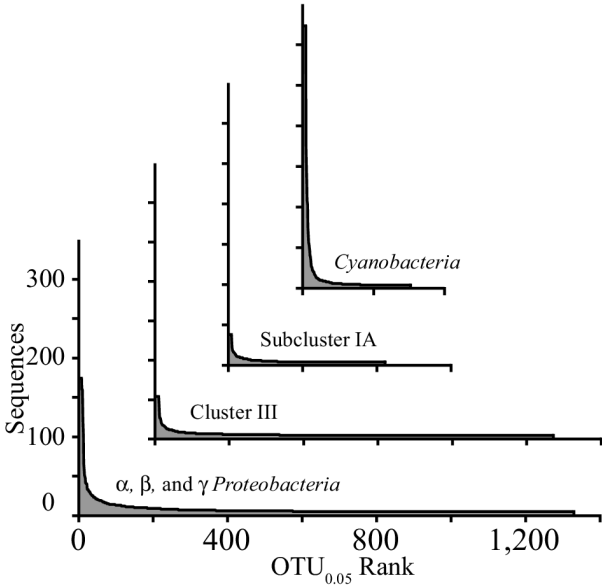


Figure 3.5: Rank-abundance distribution for $OTU_{0.05}$ within each of the major *nifH* clusters and subclusters.

documented in 14 submissions) contained the species *Katagnymene spiralis* [29], which has been recommended for incorporation into the genus *Trichodesmium* [30]. Another of the most frequently observed OTUs contained the species *T. thiebautii* (107 sequences documented in 11 submissions). The second most frequent OTU corresponds to the unicellular cyanobacterial group A (UCYN-A) (172 sequences documented in 20 studies). Another of the most frequently observed OTUs contains the species *Stenotrophomonas maltophila* (117 sequences, documented in 20 studies). Sequences from this OTU have been documented as contaminants in PCR reagents [20], and *S. maltophila* has been indicated as a common contaminant of ultrapure water systems [21].

The interaction of environment and phylogeny

The environmental affiliation of the sequences was examined with respect to phylogeny (Table 3.1). Termite guts were dominated by sequences that belonged to *nifH* cluster III (Table 3.1). Microbial mats contained primarily *nifH* sequences from cluster III and the Cyanobacteria (Table 3.1). Marine systems were dominated by *nifH* sequences from the α, β, and γ proteobacterial cluster and the *Cyanobacteria,* while soils were dominated primarily by sequences from the α, β, and γ proteobacterial cluster (Table 3.1).

Table 3.2: The 10 most frequent $OTU_{0.05}$ observed in the *nifH* database.

Rank	Seq.[a]	Source	*nifH* Cluster	Cultivated Member
1	320	marine	*Cyanobacteria*	*Katagnymene spiralis*
2	172	marine	*Cyanobacteria*	none
3	170	marine	*α*, *β*, and *γ Proteobacteria*	none
4	156	soil	*α*, *β*, and *γ Proteobacteria*	none
5	156	marine	*α*, *β*, and *γ Proteobacteria*	*Rhodobacter sphaeroides*
6	139	freshwater	*Cyanobacteria*	*Lyngbya wollei*
7	138	marine	*α*, *β*, and *γ Proteobacteria*	none
8	126	hypersaline	*Cyanobacteria*	none
9	117	PCR reagent contaminant, soil, marine	*α*, *β*, and *γ Proteobacteria*	*Stenotrophomonas maltophilia*
10	107	marine	*Cyanobacteria*	*Trichodesmium thiebautii*

[a] The number of sequences in each OTU.

Cluster III *nifH* sequences have been recovered from a wide range of environments, with 15% of all cluster III sequences observed in marine systems, 24% observed in soils, and 21% observed in microbial mats and 11% in termite guts. *Cyanobacterial nifH* sequences were common in marine, soil and mat systems with 33% of all cyanobacterial sequences observed in marine systems, 20% observed in soils, 25% observed in microbial mats and none observed in termite guts. The α, β, and γ proteobacterial cluster were most common in soil and marine environments, with 18% of all sequences from the *α*, *β*, and *γ Proteobacteria* observed in marine systems, 40% observed in soils, 2% observed in microbial mats and < 1% in termite guts. The majority of *nifH* sequences belonging to subcluster IA were observed in soils, with 11% of all subcluster IA sequences observed in marine systems, 61% observed in soils, 1% observed in microbial mats and none observed in termite guts. Subcluster IA sequences also represented 14% of all *nifH* sequences observed in soils (Table 3.1).

Discussion

The database we assembled consists of 16 989 *nifH* sequences from numerous independent studies conducted in a range of environments and sites across the globe. Accumulation curves calculated from the database suggest that while we have a near complete census of the major lineages of diazotrophs (as defined at $OTU_{0.40}$ or $OTU_{0.20}$) the discovery of new taxa (as defined at $OTU_{0.05}$) continues at a brisk pace (Figure 3.1). Nearly one out of every five taxa observed (at $OTU_{0.05}$) is represented by only a single sequence in the database generating a long tail of rare OTUs (Figure 3.4).

This pattern of distribution has been documented previously in microbial communities giving rise to the idea of the 'rare biosphere' [33]. Soils account for both the greatest number and the greatest diversity of sequences present in the database (Figure 3.2). In contrast, the marine environment is characterized by lower diversity and clear dominance by a relatively small number of nitrogen-fixing *Cyanobacteria,* most notably *Katagnymene spralis*, *T. thiebautii* and UCYN-A (Table 3.2), which represent some of the most frequently observed sequence types in the database.

The major phylogenetic clusters are not distributed equally across environments (Table 3.1). Most striking are the results for cluster III. This cluster is composed entirely of anaerobic organisms including the genera *Treponema, Spirochaeta*, *Clostridium* and *Desulfovibrio,* and various lineages of methanogens as has been described previously [34]. Termite gut communities were observed to contain many lineages from cluster III, likely relating to the anoxic nature of the termite gut [35]. Likewise, cluster III *nifH* genes were second only to *Cyanobacteria* in microbial mat communities (Table 3.1). Microbial mats are frequently home to sulfate-reducing bacteria like *Desulfovibrio* as well as fermentative spirochetes. Members of cluster III could also be observed in marine sediments and in certain soils. Most remarkable, cluster III was observed to contain 639 groups at $OTU_{0.20}$ (in contrast to only 154 for the α, β, and γ proteobacterial cluster and 47 for the *Cyanobacteria,* when 2089 sequences were sampled from each cluster). Thus, it seems clear that the highest potential for discovery of novel nitrogen-fixing organisms resides among anoxic environments. Most studies performed on nitrogen-fixing organisms to date have focused either on soils or on the photic zone of the ocean, and anoxic environments have received somewhat less attention. In addition, most universal *nifH* primer sets

have been designed using sequences from *Proteobacteria* and *Cyanobacteria.* Efforts to characterize the diversity of diazotrophs in anoxic systems would likely benefit from the creation of new primer sets designed to encompass the diversity of cluster III sequences.

While cluster III *nifH* sequences were found primarily in anoxic environments, the other clusters each demonstrated different patterns of environmental affinity. Most sequences from the *Cyanobacteria* were found in marine systems and microbial mats (Table 3.1) as would be expected [36]. *Cyanobacterial nifH* sequences were also observed in association with soils, mostly from soil crusts, photosynthetic communities that live at the soil surface in certain arid ecosystems. In contrast, most sequences from the *α*, *β*, and *γ Proteobacteria* were recovered from soils, with a large number also recovered from marine systems and a minority observed in microbial mats (Table 3.1). Sequences from subcluster IA were primarily observed in soils (Table 3.1). This cluster contains members of the *δ Proteobacteria* from the orders *Desulfuromonadales* and *Myxococcales.* Cultivated representatives include iron and sulfur-reducing anaerobic bacteria from the genera *Geobacter, Pelobacter, Desulfuromonas* and *Anaeromyxobacter.* While these cultivated representatives are found in select subgroups within subcluster IA, the majority of the subgroups within IA do not contain any cultivated representatives. Members of the IA group have been shown by $^{15}N_2$ stable isotope probing to be actively engaged in nitrogen fixation in soil [37], and this group can account for as much as half of the *nifH* genes observed in certain soils [38].

Six of the 10 $OTU_{0.05}$ that were observed most frequently in the database have been observed primarily in marine systems, and three of these are *Cyanobacteria.*

Nitrogen-fixing *Cyanobacteria* are widespread and provide an important source of N in many marine systems (as reviewed in [39]). The frequency with which these taxa are observed in the database may owe to their global distribution in marine environments, but their frequency of observation may also be a function of the number of studies conducted on nitrogen fixation in marine systems. The most abundant $OTU_{0.05}$ containing 320 sequences and observed in 14 studies includes the cultivated representative *K. spiralis*, which is closely related to *Trichodesmium* species [18, 19]. Sequences related to *T. thiebautii* are also common in the database (Table 3.2). Both of these organisms are prevalent in tropical and subtropical ocean waters. In particular, *Trichodesmium* has been shown to account for a substantial portion of primary production in these oceanic zones [27] and has been estimated to contribute 38% of nitrogen fixation in Atlantic, Pacific and Indian Ocean waters above 25°C [28]. The second most abundant OTU (Table 3.1) corresponds to the unicellular cyanobacterial group A (UCYN-A). UCYN-A was initially discovered in oligotrophic ocean waters [29], and was later shown to be both abundant and to express its nitrogenase in the subtropical North Pacific Ocean [30]. Subsequently, it has been determined that UCYN-A exhibits a low level of sequence divergence [31] and is widespread in the oceans [32-35]. UCYN-A also follows a diel pattern of nitrogenase expression where highest expression occurs during the day [34, 36, 37]. UCYN-A, although yet-to-be cultivated, has been shown through metagenomic approaches to lack oxygenic photosystem II [38] and instead has a photofermentative metabolism [31] possibly explaining the ability of the organism to fix nitrogen during the daytime.

The frequency with which sequences are observed in the database is unlikely to reflect their actual abundance in the environment, although we would expect abundant

and widespread organisms to be observed in multiple studies and for these organisms to be common in the sequence database as a result. Clearly, inter-study differences in PCR protocols and sequencing intensity provide opportunities for OTUs from certain environments to be highly overrepresented in sequence databases. Another interesting observation is that sequences affiliated with *Stenotrophomonas maltophilia* comprise one of the OTUs seen most frequently in the database (observed in 20 studies). Sequences from this OTU have been identified as a common contaminant in PCR reagents [20], and *S. maltophilia* isolates have been documented as contaminants in ultrapure water systems [21]. This finding suggests that the frequency with which this OTU has been observed may have more to do with frequency of finding DNA from this organism in the laboratory environment than the frequency with which this organism occurs in natural environments.

We have undertaken the first comprehensive assessment of *nifH* richness as a function of environment and phylogenetic affiliation. Our findings reveal that much diversity still awaits discovery, particularly among anaerobic nitrogen fixers and in the soil environment. The database we have created is a resource that may be used for further exploration of the sequence data as well as to develop and evaluate PCR primers to target undersampled phylogenetic groups and environments.

Experimental procedures

Constructing the *nifH* database

Construction of the *nifH* database began by downloading all sequence records

containing a *nifH* related entry from the GenBank Nucleotide Database [58]. A simple search for the term *nifH* was insufficient to recover all *nifH* sequences, and the search query also included the terms dinitrogenase reductase and nitrogenase iron protein in various arrangements. These records were manually vetted to eliminate non-*nifH* sequences caught by the search. Additionally, *nifH* sequences were extracted from genomes and multi-CDS records. The *nifH* sequences were imported into ARB [40] along with a seed alignment used to guide the sequence alignment process. The seed alignment was based on the Fer4_NifH (PF00142) protein family alignment obtained from Pfam [41]. The Fer4_NifH seed alignment was reverse-translated into nucleotide sequences using BioEdit [42]. The Fer4_NifH seed alignment was used to construct a PT_server in ARB, and this PT_server was used to align the *nifH* records imported from GenBank. The reverse-translated sequences of the Pfam starter alignment were then deleted from the database. The alignments were visually inspected and manually corrected. Poorly aligned or difficult-to-align sequences were detected through phylogenetic analysis in later stages of database construction and were subsequently removed. The database contains all *nifH* sequences submitted to GenBank until 2/4/2009 and is available for download in ARB database format at http://www.css.cornell.edu/faculty/ buckley/nifH_database_2_4_09.arb.

Richness estimates

The diversity of nifH sequences in the database was evaluated using DOTUR [62] to cluster sequences based on nucleic acid sequence similarity into OTUs, and the cut-offs used herein are represented with the convention $OTU_{0.05}$, in which the OTU

cut-off criterion is provided in subscript. The $OTU_{0.05}$ nucleotide sequence cutoff for conserved protein encoding sequences is expected to correspond very roughly to the level of microbial species [44]. The *nifH* sequence fragments in the database vary in length and position in the gene alignment complicating efforts to make a single distance matrix that includes all sequences. To solve this problem, we identified a portion of the sequence alignment that provided the greatest number of overlapping sequences of sufficient length for substantive analysis. The region identified for DOTUR analyses spanned the nucleotide positions 133 to 454 (numbering based on the *nifH* gene sequence of *Azotobacter vinelandii*, ACCN# M20568). As a result, a total of 10 833 sequences were ultimately used in DOTUR analyses (paralogous sequences that belong to clusters IV and V were excluded from DOTUR analyses and are not included in this total). The richness estimates obtained using positions 133 to 454 matches closely with estimates made with other regions of the *nifH* gene as determined by analyses of the Chao1 richness estimate for different regions of the gene sequence. The distance matrix was calculated in ARB and exported for use in DOTUR analyses. DOTUR's default, furthest neighbour clustering method was used as well as randomized input order and rarefaction. Pielou's J, or evenness, was calculated from the DOTUR output by dividing the final sampling value of the Shannon diversity index for each of the phylogenetic clusters by the natural logarithm of the number of species sampled.

The association of sequences with different environmental categories was achieved by searching for environment associated keywords in all the fields of the GenBank entries (Table 3.1). Sequences of each type were marked and then manually verified. Sequences that could not be associated with an environment were excluded

from the environment analysis. Search terms were combined in logical expressions for each query to recover relevant sequence entries while excluding confounding entries. For example, sequence entries for the marine category included sequences from coastal and open ocean water column samples but excluded marine sediments, mats, estuaries and wetlands. The soils category includes rhizosphere and bulk soil samples from terrestrial habitats but excludes soils in wetlands and estuaries. Further details about the environmental sources of sequences in these categories can be found by examining the saved configurations of these categories in the ARB *nifH* database. Whereas the different environmental and phylogenetic categories that we evaluated each contain different numbers of sequences, and the Chao1 richness estimator is influenced strongly by the number of sequences sampled, we controlled for sampling intensity in categorical comparisons by using the lowest number of sequences common to each category when calculating Chao1 estimates. As described above, sequences used to calculate Chao1 were drawn randomly without replacement from the set of sequences in each category.

Phylogenetic analyses

Phylogenetic analyses were complicated by the presence of non-overlapping sequence fragments in the database, as described in the above section. Two approaches were used to construct phylogenetic trees and to determine the phylogenetic affiliation of individual sequences. Initially, a comprehensive guide tree was created in ARB to facilitate database management and selection of sequences. The guide tree was created by first using neighbour joining to construct a tree from a non-redundant set of 6878

sequences that had sequence information between nucleotide positions 133 to 454. Additional sequences were added to the tree in batches using the quick add by parsimony function in ARB. The final tree contained 16 989 sequences. The guide tree was used primarily for database navigational purposes.

We based selection of phylogenetic clusters upon several considerations. First, wherever possible we identified clusters in a manner consistent with affiliations identified in Zehr and colleagues [64]. Second, in order to identify objective criteria for defining phylogenetic clusters, we examined OTUs formed at different levels of sequence similarity and found that an $OTU_{0.40}$ cut-off roughly corresponded to the major distinct monophyletic lineages we observed in phylogenetic analyses. This $OTU_{0.40}$ criterion generated subclusters within cluster I that corresponded to the α, β, and γ *Proteobacteria*, the *Cyanobacteria* and subcluster IA. Subcluster IA was first described by Zehr and colleagues [65] at which time it was composed of 65 sequences and contained no cultivated isolates. Our analyses show that this cluster contains δ *Proteobacteria* of the *Desulfuromonadales* and *Myxococcales*. The $OTU_{0.40}$ criterion, however, yielded a large number of subclusters when applied to cluster III, the membership of these subclusters in cluster III was too small to allow robust independent analysis. Thus, we chose to maintain cluster III as a single entity in our analyses as it has been described by previous groups [66][7][8]. Finally, we did not perform analyses of diversity within cluster II, or within several $OTU_{0.40}$ groups that are present at the base of the *Cyanobacteria* (containing *Frankia*, *Paenibacillus* and ε *Proteobacteria*) because these groups contained too few sequences to permit robust analysis.

References

1. Vitousek PM, Howarth RW (1991) Nitrogen limitation on land and in the sea: how can it occur? Biogeochemistry 13: 87-115.

2. Rubio LM, Ludden PW (2002) The gene products of the *nif* regulon. In: Leigh GJ, editor. Nitrogen fixation at the millennium. Elsevier Science B.V. pp. 101-136.

3. Zehr JP, McReynolds LA (1989) Use of degenerate oligonucleotides for amplification of the *nifH* gene from the marine cyanobacterium Trichodesmium thiebautii. Appl Environ Microbiol 55: 2522-2526.

4. Zehr JP, Jenkins BD, Short SM, Steward GF (2003) Nitrogenase gene diversity and microbial community structure: a cross-system comparison. Environ Microbiol 5: 539-554.

5. Young JPW (1992) Phylogenetic classification of nitrogen-fixing organisms. In: Stacey G, Burris RH, Evans HJ, editors. Biological nitrogen fixation. Chapman and Hall. pp. 43-86.

6. Chien Y, Zinder SH (1994) Cloning, DNA-sequencing, and characterization of a *nifD*-homologous gene from the archaeon *Methanosarcina barkeri* 227 which resembles *nifD1* from the eubacterium *Clostridium pasteurianum*. J Bacteriol 176: 6590-6598.

7. Raymond J, Siefert JL, Staples CR, Blankenship RE (2004) The natural history of nitrogen fixation. Mol Biol Evol 21: 541-554.

8. Young JPW (2005) The phylogeny and evolution of nitrogenases. In: Palacios R, Newton WE, editors. Genomes and genomics of nitrogen-fixing organisms. Springer. pp. 221-241.

9. Bürgmann H, Widmer F, Von Sigler W, Zeyer J (2004) New molecular screening tools for analysis of free-living diazotrophs in soil. Appl Environ Microbiol 70: 240-247.

10. Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol 62: 625-630.

11. Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K et al. (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiol

Ecol 60: 341-350.

12. Ueda T, Suga Y, Yahiro N, Matsuguchi T (1995) Remarkable N_2-fixing bacterial diversity detected in rice roots by molecular evolutionary analysis of *nifH* gene sequences. J Bacteriol 177: 1414-1417.

13. Marusina AI, Boulygina ES, Kuznetsov BB, Tourova TP, Kravchenko IK et al. (2001) A system of oligonucleotide primers for the amplification of *nifH* genes of different taxonomic groups of prokaryotes. Mikrobiologiya 70: 86-91.

14. Poly F, Monrozier LJ, Bally R (2001) Improvement in the RFLP procedure for studying the diversity of *nifH* genes in communities of nitrogen fixers in soil. Res Microbiol 152: 95-103.

15. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. Appl Environ Microbiol 67: 4399-4406.

16. Souillard N, Magot M, Possot O, Sibold L (1988) Nucleotide sequence of regions homologous to *nifH* (nitrogenase Fe protein) from the nitrogen-fixing archaebacteria *Methanococcus thermolithotrophicus* and *Methanobacterium ivanovii*: evolutionary implications. J Mol Evol 27: 65-76.

17. Staples CR, Lahiri S, Raymond J, Von Herbulis L, Mukhophadhyay B et al. (2007) Expression and association of group IV nitrogenase NifD and NifH homologs in the non-nitrogen-fixing archaeon *Methanocaldococcus jannaschii*. J Bacteriol 189: 7392-7398.

18. Lundgren P, Söderbäck E, Singer A, Carpenter EJ, Bergman B (2001) *Katagnymene*: Characterization of a novel marine diazotroph. J Phycol 37: 1052-1062.

19. Orcutt KM, Rasmussen U, Webb EA, Waterbury JB, Gundersen K et al. (2002) Characterization of *Trichodesmium* spp. by genetic techniques. Appl Environ Microbiol 68: 2236-2245.

20. Zehr JP, Crumbliss LL, Church MJ, Omoregie EO, Jenkins BD (2003) Nitrogenase genes in PCR and RT-PCR reagents: implications for studies of diversity of functional genes. Biotechniques 35: 996-1002, 1004-1005.

21. Kulakov LA, McAlister MB, Ogden KL, Larkin MJ, O'Hanlon JF (2002) Analysis of bacteria contaminating ultrapure water in industrial systems. Appl Environ Microbiol 68: 1548-1555.

22. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103: 12115-12120.

23. Brune A, Friedrich M (2000) Microecology of the termite gut: structure and function on a microscale. Curr Opin Microbiol 3: 263-269.

24. Buckley DH, Huangyutitham V, Hsu S, Nelson TA (2007) Stable isotope probing with $^{15}N_2$ reveals novel noncultivated diazotrophs in soil. Appl Environ Microbiol 73: 3196-3204.

25. Hsu S, Buckley DH (2009) Evidence for the functional significance of diazotroph community structure in soil. ISME J 3: 124-136.

26. Paerl HW, Zehr JP (2000) Microbial ecology of the oceans. In: Kirchman DL, editor. Wiley-Liss. pp. 387-426.

27. Karl D, Michaels A, Bergman B, Capone D, Carpenter E et al. (2002) Dinitrogen fixation in the world's oceans. Biogeochemistry 57: 47-98.

28. Mahaffey C, Michaels AF, Capone DG (2005) The conundrum of marine N_2 fixation. Am J Sci 305: 546-595.

29. Zehr JP, Mellon MT, Zani S (1998) New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of Nitrogenase *(nifH)* genes. Appl Environ Microbiol 64: 3444-3450.

30. Zehr JP, Waterbury JB, Turner PJ, Montoya JP, Omoregie E et al. (2001) Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean. Nature 412: 635-638.

31. Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA et al. (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. Nature 464: 90-94.

32. Langlois RJ, LaRoche J, Raab PA (2005) Diazotrophic diversity and distribution in the tropical and subtropical Atlantic ocean. Appl Environ Microbiol 71: 7910-7919.

33. Langlois RJ, Hümmer D, LaRoche J (2008) Abundances and distributions of the dominant *nifH* phylotypes in the Northern Atlantic Ocean. Appl Environ Microbiol 74: 1922-1931.

34. Needoba JA, Foster RA, Sakamoto C, Zehr JP, Johnson KS (2007) Nitrogen

fixation by unicellular diazotrophic cyanobacteria in the temperate oligotrophic North Pacific Ocean. Limnology and Oceanography 52: 1317-1327.

35. Moisander PH, Beinart RA, Hewson I, White AE, Johnson KS et al. (2010) Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. Science 327: 1512-1514.

36. Church MJ, Short CM, Jenkins BD, Karl DM, Zehr JP (2005) Temporal patterns of nitrogenase gene *(nifH)* expression in the oligotrophic North Pacific Ocean. Appl Environ Microbiol 71: 5362-5370.

37. Zehr JP, Montoya JP, Jenkins BD, Hewson I, Mondragon E et al. (2007) Experiments linking nitrogenase gene expression to nitrogen fixation in the North Pacific subtropical gyre. Limnology and Oceanography 52: 169-183.

38. Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F et al. (2008) Globally distributed uncultivated oceanic N_2-fixing cyanobacteria lack oxygenic photosystem II. Science 322: 1110-1112.

39. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. Nucleic Acids Res 36: database issue D25-D30.

40. Ludwig W, Strunk O, Westram R, Richter L, Meier H et al. (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32: 1363-1371.

41. Finn RD, Mistry J, Tate J, Coggill P, Heger A et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211-22.

42. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41: 95-98.

43. Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol 71: 1501-1506.

44. Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci 361: 1929-1940.

CHAPTER 4

A COMPREHENSIVE EVALUATION OF PCR PRIMERS TO AMPLIFY THE *NIFH* GENE OF NITROGENASE

Introduction

Nitrogen-fixing microorganisms are globally significant in that they provide the only natural biological source of fixed nitrogen in the biosphere. These organisms enzymatically transform dinitrogen gas from the atmosphere into ammonium equivalents needed for biosynthesis of essential cellular macromolecules. Nitrogen-fixing bacteria are diverse, and most of the known taxa have not yet been cultivated in the laboratory [1]. Nitrogen fixation is carried out by the nitrogenase enzyme whose multiple subunits are encoded by the genes *nifH*, *nifD*, and *nifK* (as reviewed in [2]). Of the three, *nifH* (encoding the nitrogenase reductase subunit) is the most sequenced and has become the marker gene of choice for researchers studying the phylogeny, diversity, and abundance of nitrogen-fixing microorganisms. Thus, many PCR primers have been developed to target the *nifH* gene with the purpose of amplifying this gene sequence from environmental samples.

Through use of *nifH* as a marker gene, researchers have been able to characterize aspects of the diversity and ecology of nitrogen-fixing *Bacteria* and *Archaea*. A wide range of environments have been sampled for *nifH* gene diversity

including marine [3], terrestrial [4], extreme [5], anthropogenic [6], host-associated [7], and agricultural [8]. Analysis of these data indicate that the distribution of diazotrophs in the environment varies as a function of habitat type [1]. While more than 3,358 $OTU_{0.05}$ *nifH* sequence types have been determined, the global census of diazotroph diversity remains far from complete [9]. Rates of nitrogen fixation have been associated with both *nifH* abundance [10] and *nifH* diversity [11], and thus knowledge of diazotroph community structure and dynamics is required to understand the ecological constraints on nitrogen fixation in microbial communities.

Phylogenetic analyses of *nifH* gene sequences have revealed five primary clusters of genes homologous to *nifH* [12-15]. Cluster I consists of aerobic nitrogen fixers including *Proteobacteria*, *Cyanobacteria*, *Frankia*, and *Paenibacillus*. Cluster II is generally thought of as the alternative nitrogenase cluster because it contains sequences from FeFe and FeV nitrogenases which differ from the conventional FeMo cofactor-containing nitrogenase. Cluster III consists of anaerobic nitrogen fixers from *Bacteria* and *Archaea* including for instance the *Desulfovibrionaceae*, *Clostridium*, *Spirochataes*, and *Methanobacteria*. Cluster IV and cluster V contain sequences that are paralogs of *nifH* and which are not involved in nitrogen fixation [13].

We set out to provide a comprehensive evaluation of primer coverage for researchers wishing to use the *nifH* gene as a molecular marker for the study of nitrogen-fixing *Bacteria* and *Archaea*. Primers that target diverse *nifH* sequences must be degenerate to encompass the sequence variability of the *nifH* gene, and Zehr and

McReynolds were the first to design such degenerate primers [16, 17]. There have since been numerous efforts to design both universal and group-specific *nifH* primer sets. In a survey of the literature, we have found 51 universal and 35 group-specific primers that have been paired to make 42 universal and 19 group-specific primer sets. We have performed an *in silico* evaluation of all of these *nifH* primers using an aligned database of all publicly available *nifH* sequences which we constructed previously [9]. We then performed empirical tests of the best of these primers using genomic DNA from a phylogenetically diverse set of nitrogen fixers and DNA from soil.

Results

Any effort to assess PCR primer coverage *in silico* must account for variation in sequence depth along the gene alignment of the database being queried. We observe that nucleotide positions near the beginning and end of the *nifH* gene alignment are under-represented in sequence databases relative to nucleotide positions in the middle of the gene alignment (Figure 4.1). This problem occurs because a majority of *nifH* sequences have been generated using PCR primers that bind to conserved nucleotide positions found within the *nifH* gene. A majority of the 393 full-length *nifH* sequences currently present in the *nifH* database are derived from sequenced genomes. The two dips in nucleotide coverage (at position 199 and 350 in Figure 4.1) result from insertions in the *Azotobacter vinelandii nifH* reference sequence relative to other genes in the alignment. In addition, some sequences in the alignment have insertions relative to *A. vinelandii* (data not shown). Due to the variations observed in sequence depth

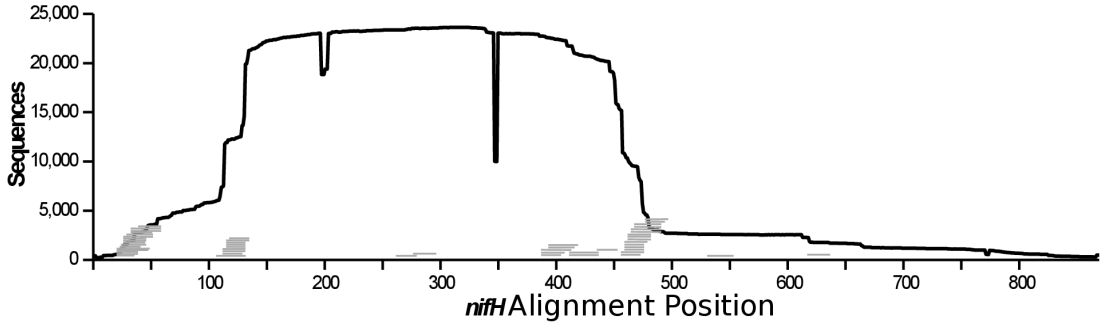


Figure 4.1: Coverage of the *nifH* gene by sequences and primers in the *nifH* database. The number of sequences in the *nifH* database is depicted in relation to alignment position along the gene. Alignment positions are referenced to the *nifH* nucleotide position from *Azotobacter vinelandii* (Genbank ACCN# M20568). Universal *nifH* primer sequences listed in Table 4.1 are indicated by grey horizontal lines.

along the alignment, all estimates of primer coverage were calculated with respect to the total number of sequences available at the alignment positions where each primer binds.

We mapped the 51 universal primers to their complementary binding positions along the *A. vinelandii nifH* gene (Figure 4.1). Many primers bind to the same region (Figure 4.1), and thus may vary only slightly in binding position, oligonucleotide length, or degeneracy.

The quality and characteristics of universal *nifH* PCR primers vary widely (Table 4.1). Of the universal primers 15 of the 51 were found to hit 90% or more of all *nifH* sequences while 23 hit less than 50% of these sequences and 9 hit 10% or fewer sequences (Table 4.1). In general, those universal primers that had > 90% coverage for clusters I and III did not demonstrate systematic bias against individual phylogenetic groups within these clusters (Table 4.1). The primer KAD3 is notable, however,

Table 4.1. Properties of universal primers and their coverage for phylogenetic and environmental groupings in the *nifH* database.

Table 4.1

Primer [d]	Name	Pos. [e]	Deg. [f]	T_m (°C)	*nifH* [a] (%)			Specific groupings [b] (%)								Environ. [c] (%)			Ref. [g]
					0	1	2	Pr	Cy	III	IA	Fr	Pb	Ep	IV	Soil	Mat	Sea	
GCIWTYTAYGGNAARGG	Nh21F	19-35	64	51.8-61.9	93	96	103	89	100	100	100	100	91	83	76	78	100	100	[42]
GCIWTITAYGGNAARGGNGG	nifH19F	19-38	128	59.5-69.5	94	96	96	89	100	100	100	100	91	100	79	78	100	100	[43]
GCIWTYTAYGGIAARGGIGG	Ueda19F	19-38	16	62.4-67.9	93	96	96	89	100	100	100	100	91	83	76	78	100	100	[44]
GCIWTHTAYGGIAARGGIGGIATHGGIAA	IGK3	19-47	72	69.4-75.3	92	95	95	87	100	98	96	100	91	100	78	72	100	100	[45]
GCGTTCTACGGTAAGGGCGGTATCGGNAAR	K07-F	19-48	8	71.0-72.8	1	3	10	2	0	0	0	0	0	0	0	0	0	0	[27]
TTYTAYGGNAARGGNGG	nifH4	22-38	128	49.8-63.5	52	91	94	71	100	4	43	93	91	13	23	43	100	78	[46]
TCTACGGAAAGGGCGGTATCGG	primer-f	23-44	1	66.5	10	14	31	13	0	13	0	0	0	23	0	0	0	0	[47]
TACGGCAARGGTGGNATHG	FGPH19	25-43	24	58.2-66.2	7	24	63	9	0	4	1	71	0	2	4	3	0	0	[48]
TACGGYAARGGBGGYATCGG	IGK-Poly [h]	25-44	24	60.3-70.6	18	50	72	22	45	6	16	29	17	0	9	6	100	33	[20]
TACGG(P/K)AAKGG(P/G)GG(P/K)ATPGG	PicenoF44	25-44	8	NA	50	85	96	59	96	35	25	79	33	59	31	34	100	89	[49]
TAYGGIAARGGIGGIATYGGIAARTC	F1	25-50	4096	60.4-74.5	79	95	96	84	100	69	65	93	83	87	80	83	100	100	[18]
GGHAARGGHGGHATHGGNAARTC	MehtaF	28-50	1296	57.4-72.9	59	78	87	46	96	77	44	94	44	51	69	67	100	100	[5]
AARGGNGGNATHGGNAA [i]	IGK	31-47	384	62.1-72.5	90	98	104	91	100	93	70	95	89	94	95	95	100	100	[26]
AAAGGYGGWATCGGYAARTCCACCAC	nifHF-Rösch [h]	31-56	16	66.0-71.6	14	25	44	11	34	5	9	70	22	14	0	34	100	5	[24]
AAAGGYGGWATCCGYAARTCCACCAC	rösch F-1b [h]	31-56	16	66.0-71.6	0	14	25	0	0	0	0	0	0	0	0	0	0	0	[25]
GGTATYGGYAARTCSACSAC	RL28	37-56	32	57.7-64.1	41	70	84	43	21	41	62	57	6	12	3	24	100	5	[50]
ATHGTIGGITGYGAYCCIAARGCIGA	KAD3	106-131	16	70.1-76.8	70	84	94	89	90	15	80	97	80	82	1	70	100	24	[45]
GGNTGYGAYCCNAARGC	469	112-128	128	53.4-67.4	88	92	98	82	94	92	95	98	79	96	35	77	100	73	[20]
GGITGTGAYCCNAAVGCNGA	nif112	112-131	96	60.9-70.4	29	90	92	21	37	30	34	36	28	69	11	30	60	22	[43]
GGITGYGAYCCNAAVGCNGA	nifH-univ-f112	112-131	192	60.9-72.7	88	91	98	82	91	92	96	98	79	97	35	77	100	73	[21]
TGYGAYCCNAARGCNGA	nifH2	115-131	128	54.0-68.1	95	98	98	93	98	96	98	98	98	96	37	93	99	89	[16]
TGYGAYCCIAARGCIGA	Kadino	115-131	8	60.2-67.9	95	98	98	93	98	96	98	98	98	96	37	93	99	89	[20]
TGYGAYCCIAAIGCIGA	F2	115-131	4	62.3-67.9	96	98	98	94	98	98	99	99	98	98	37	94	99	93	[18]
TGCGAYCCSAARGCBGACTC	polF	115-134	24	63.8-70.1	39	70	88	51	12	26	41	61	34	6	3	54	15	8	[20]
GAYCCNAARGCNGACTC	nifH11	118-134	64	52.7-63.4	72	96	98	79	67	50	82	97	78	48	17	78	66	55	[51]
CTCCGGGCCRCCNGAYTC	FGPH273'	262-279	16	63.7-70.7	11	44	79	11	0	3	6	93	0	3	1	14	2	5	[48]
GMRCCIGGIGTIGGYTGYGC	nifH-2f	277-296	16	69.2-78.3	87	99	99	92	80	74	94	98	95	72	20	85	85	79	[19]
CCRCCRCANACMACGTC	Cy55Nh428R	388-404	32	56.6-67.5	66	94	99	68	76	49	63	76	55	49	23	56	65	80	[42]
AAICCRCCRCAIACIACRTC	Ueda407R	388-407	8	63.9-70.6	91	99	99	92	95	85	90	99	64	79	71	87	95	93	[44]
ATIGCRAAICCICCRCAIACIACRTC	DVV	388-413	8	71.7-75.8	94	98	99	93	94	94	96	98	90	95	52	91	96	93	[45]
GGCATNGCRAANCCVCCRCANAC	MehtaR	394-416	768	63.2-75.1	92	98	99	92	94	89	95	98	89	89	49	89	97	92	[5]
ATIGGCATIGCRAAICCICCRCAIAC	VCG	394-419	4	73.9-76.7	93	98	99	93	94	94	95	98	90	96	33	91	96	91	[45]
TGGGCYTTGTTYTCRCGGATYGGCAT	nifHRc	412-437	16	69.1-74.2	11	34	51	18	0	0	11	0	13	0	0	12	3	9	[24]
TGSGCYTTGTCYTCRCGGATBGGCAT	nifHRb	412-437	48	70.0-76.0	0	33	56	1	0	0	0	0	0	0	0	1	0	0	[24]
SACGATGTAGATPTCCTG	PicenoR436	436-453	4	NA	34	60	83	51	10	24	13	97	7	1	3	36	28	22	[49]
TCIGGIGARATGATGGC	R6	457-473	2	61.1-62.5	96	97	99	97	99	86	99	99	102	97	15	94	99	97	[18]
ATSGCCATCATYTCRCCGGA	polR	457-476	8	63.7-67.5	35	63	86	36	15	19	40	90	23	12	0	55	7	2	[20]
ADNGCCATCATYTCNCC	nifH1	460-476	96	52.5-63.9	94	99	99	94	96	91	96	99	103	80	13	91	98	87	[16]
ADWGCCATCATYTCRCC	nifH22	460-476	24	53.2-60.9	17	89	98	16	23	21	22	1	49	20	3	11	31	36	[51]

Table 4.1 continued.

Primer [d]	Name	Pos. [e]	Deg. [f]	T_m (°C)	*nifH* [a] (%)			Specific groupings [b] (%)								Environ. [c] (%)			Ref. [g]
					0	1	2	Pr	Cy	III	IA	Fr	Pb	Ep	IV	Soil	Mat	Sea	
ANDGCCATCATYTCNCC	nifH2-ZANI [i]	460-476	96	52.5-63.6	54	98	99	48	63	70	73	11	83	83	10	63	61	76	[46]
TANANNGCCATCATYTCNCC	470	460-479	512	53.8-65.7	80	82	98	84	43	62	88	99	80	98	9	79	6	95	[20]
GCRTAIABNGCCATCATYTC	nifH-univ-463r	463-482	48	55.7-63.8	85	87	88	91	50	62	91	99	88	99	8	93	83	72	[43]
GCRTAIAIIGCCATCATYTC	Emino	463-482	4	60.2-63.4	86	87	88	91	50	64	91	99	88	100	8	93	83	76	[20]
ATGATGGCSATGTAYGCSGCSAACAA	nifHR-2 [i]	466-491	16	70.0-71.7	49	58	87	36	17	35	53	99	60	58	0	72	100	4	[24]
TTGTTSGCSGCRTACATSGCCATCAT	nifHR	466-491	16	70.0-71.7	49	58	87	36	17	35	53	99	60	58	0	72	100	4	[27]
TTGTTGGCIGCRTASAKIGCCAT	nifH-3r	469-491	8	68.5-72.1	48	89	94	66	7	39	72	4	30	44	3	77	100	0	[19]
ATRTTRTTNGCNGCRTA	nifH3	494-478	128	46.1-61.5	94	95	98	93	96	86	88	100	78	93	50	93	100	100	[46]
YAAATRTTRTTNGCNGCRTA	YAA-poly [i]	478-497	256	49.5-63.5	1	12	51	0	0	7	1	0	0	0	5	0	0	21	[20]
CAGATCAGVCCGCCSAGRCGMAC	RL25	532-554	24	67.5-74.1	4	29	61	5	0	8	1	0	0	0	0	1	0	0	[50]
GGCACGAAGTGGATCAGCTG	primer-r	619-638	1	64.3	4	16	43	3	0	24	0	0	0	29	0	0	0	0	[47]
GCTACTACYTCGCCSGA	AMR-R				0	0	0	0	0	0	0	0	0	0	0	0	0	0	[27]

[a] Data indicate primer binding to all *nifH* sequences in the database with 0, 1, and 2 mismatches allowed. In some cases highly degenerate primers bind to multiple positions in the sequence generating coverage values that exceed 100%.

[b] Data indicate primer binding to specific groupings in the *nifH* phylogeny. Abbreviations are as follows: *Alpha, Beta, and Gamma Proteobacteria* (**Pr**); *Cyanobacteria* (**Cy**)**;** Cluster III (**III**); Cluster IA (**IA**); *Paenibacillus* (**Pb**); *Frankia* (**Fr**); *Epsilon Proteobacteria* Containing Cluster (**Ep**); paralogous sequences in Cluster IV (**IV**).

[c] Primer coverage queried against sequences recovered from specific environments (Environ.) as described in methods. Environments include: soils (**Soil**), microbial mats (**Mat**), and pelagic marine samples (**Sea**).

[d] Sequences are given in the 5' to 3' direction, IUPAC characters are used, and I=Inosine.

[e] Position is relative to *A. vinelandii nifH* (Genbank ACCN# M20568).

[f] Degeneracy is given as the number of oligonucleotides that comprise the primer.

[g] References in which the primers are described.

[h] We altered these primer names in order to distinguish them from primers with similar name and sequence composition that originate from other sources.

NA: Data not available as described in Methods.

because it misses much of cluster III relative to cluster I (Table 4.1). Those primers with the highest coverage also tended to recognize a number of non-target sequences from cluster IV (Table 4.1).

The group-specific primers we evaluated generally show poor coverage of the phylogenetic groups they have been designed to target, except for the *Frankia*-specific primers nifH-f1-forA, nifH-f1-forB, nifH-269, and nifH-f1-rev (Table 4.2). The primer cyanoR targets *Cyanobacteria*, but has coverage of only 25%, and its intended pair, primer cyanoF, has a coverage of only 1% of cyanobacterial sequences (Table 4.2). Given that PCR requires two primers used in combination, a useful indication of specificity must account for the coverage obtained when using specific primer pairs (Tables 4.3 and 4.4). We evaluated both primer combinations that have been reported in the literature as well as new primer combinations. As expected, the coverage obtained with primer pairs is always lower than the coverage obtained for each individual primer. We evaluated 42 universal primer pair combinations, of which 7 hit >90% of *nifH* sequences in the database, 24 hit >50%, and 6 hit 10% or less. Those primer sets which had >90% coverage are 19F/nifH3, Nh21F/nifH1, Nh21F/nifH3, IGK/nifH3, F2/R6, nifH2/R6, and nifH1/nifH2 (ie: the Zehr and McReynolds primers). The 6 primer sets which hit 10% or less of cluster I and III are Primer-f/Primer-r, FGPH19/FGPH273', FGPH19/PolR, IGK/FGPH273', nifHF/nifHRb, and nifHF/nifHRc. While we evaluated 19 group-specific primer combinations, very few primer sets had high coverage of the designated target groups (Table 4.4). The primer

Table 4.2 Properties of group-specific primers and their coverage for phylogenetic and environmental groupings in the *nifH* database.

Table 4.2

Primer [d]	Name	Tg.[e]	Pos.[f]	Deg.[g]	T_m (°C)	*nifH* [a] (%)			Specific groupings [b] (%)								Environ. [c] (%)			Ref. [h]
						0	1	2	*Pr*	*Cy*	III	IA	*Fr*	*Pb*	Ep	IV	Soil	Mat	Sea	
CGCIWTYTACGGIAARGGIGG	ChenBR1	BR	18-38	512	66.6-69.8	42	81	94	63	6	10	60	85	60	27	9	0	0	0	[52]
GCSTTCTACGGMAAGGGTGG	nifH-f1-forA	Fr	19-38	4	63.9-66.7	3	7	24	0	0	0	0	85	0	0	0	0	0	0	[21]
GCRTTYTACGGYAARGGSGG	nifH-a1-forA	AP	19-38	32	60.6-69.1	14	38	70	26	20	0	4	0	18	0	0	0	100	11	[21]
TACGGNAARGGSGGNATCGGCAA	nifHF	R	25-47	64	66.7-73.9	20	47	78	32	4	11	12	0	17	7	10	17	100	11	[53]
GGTATYGGYAARTGYACYAC	primer-3	RA	37-56	32	52.6-64.8	0	19	56	0	0	0	0	0	0	0	0	0	0	0	[54]
GGCAAGTCCACCACCCAGC	nifHf1	Fr	43-61	1	67.0	1	2	4	0	0	0	0	30	0	0	0	0	0	0	[55]
ATYGTCGGYTGYGAYCCSAARGC	Olsen1	AM	106-128	64	65.0-73.6	37	66	81	53	2	3	58	57	0	14	0	38	100	0	[56]
CGTAGGTTGCGACCCTAAGGCTGA	cyanoF	Cy	108-131	1	68.8	0	0	1	0	1	0	0	0	0	0	0	0	0	0	[56]
GGCTGCGATCCCAAGGCTGA	nifH-b1-forB	AB	112-131	1	68.3	1	10	32	2	0	0	1	0	0	0	0	2	0	0	[21]
GGTTGTGACCCGAAAGCTGA	nifH-g1-forB	GP	112-131	1	64.1	0	3	10	1	1	0	0	0	0	0	0	1	0	0	[21]
GGWTGTGATCCWAARGCVGA	nifH-c1-forB	AN	112-131	24	58.7-64.3	1	8	25	1	1	1	0	0	0	4	0	1	0	4	[21]
GGCTGCGATCCGAAGGCCGA	nifH-a2-forB	AP	112-131	1	70.3	1	10	33	2	0	0	0	0	23	0	0	1	0	0	[21]
GGMTGCGAYCCSAARGCSGA	nifH-a1-forB	AP	112-131	32	66.2-72.7	27	58	73	29	1	31	41	22	33	5	8	15	20	0	[21]
GGBTGYGACCCSAASGCYGA	nifH-f1-forB	Fr	112-131	48	65.9-72.9	22	48	74	15	1	9	17	91	2	3	3	22	20	0	[21]
ACCCGCCTGATCCTGCACGCCAAGG	nifHFor	MS	136-160	1	74.7	11	20	32	21	0	0	5	0	0	0	0	16	0	9	[57]
TAARGCTCAAACTACCGTAT	cylnif-F	Cs	156-175	2	56.2-57.9	1	1	3	0	3	0	0	0	0	0	0	0	0	1	[58]
GAAGGTCGGCTACCAGAACA	NIFH2F	TB	231-250	1	63.1	0	2	6	1	0	0	0	0	0	0	0	2	0	0	[59]
AAGTTGATCGAGGTGATGACG	NIFH5R	TB	306-326	1	61.6	10	22	36	21	0	0	0	0	0	0	0	18	0	7	[59]
CCGGCCTCCTCCAGGTA	nifH-269	Fr	325-341	1	64.2	3	3	3	0	0	0	0	85	0	0	0	4	0	0	[60]
ATTTAGACTTCGTTTCCTAC	cylnif-R	Cs	356-375	1	54.6	1	1	4	0	3	0	0	0	0	0	0	0	0	1	[58]
ACGATGTAGATTTCCTGGGCCTTGTT	NifHRev	MS	427-452	1	67.5	13	29	43	23	0	0	6	0	3	0	0	15	1	3	[57]
GACGATGTAGATYTCCTG	primer 4=AQE	RA	436-453	2	53.8-55.1	24	54	81	33	8	21	12	69	7	0	1	19	24	13	[54]
GCATACATCGCCATCATTTCACC	cyanoR	Cy	460-482	1	63.6	4	8	23	2	25	0	1	0	0	0	0	1	0	8	[56]
GCGTACATSGCCATCATCTC	nifH-f1-rev	Fr	463-482	2	62.2-62.3	23	44	60	14	0	3	35	94	0	7	0	20	0	0	[21]
GCGTACATGGCCATCATCTC	nifH-b1-rev	AB	463-482	1	62.3	8	32	53	9	0	3	33	6	0	7	0	18	0	0	[21]
GCGTACATGGCCATCATCTC	nifH-g1-rev	GP	463-482	1	62.3	8	32	53	9	0	3	33	6	0	7	0	18	0	0	[21]
GCATAYASKSCCATCATYTC	nifH-c1-rev	AN	463-482	8	55.4-62.3	1	13	58	1	0	3	1	0	0	4	0	2	0	0	[21]
GCGTAGAGCGCCATCATCTC	nifH-a2-rev	AP	463-482	1	64.0	2	17	43	4	0	0	1	0	0	1	0	3	0	0	[21]
GCATAGAGCGCCATCATCTC	nifH-a1-rev	AP	463-482	1	62.0	9	16	33	17	0	1	0	0	0	0	0	2	0	0	[21]
ATGGTGTTGGCGGCRTAVAKSGCCATCAT	Olsen2	AM	466-494	24	71.5-75.3	0	32	54	0	0	0	0	0	0	0	0	0	0	0	[56]
CTCGATGACGGTCATCCGGC	nifHr	Fr	671-690	1	65.9	0	3	6	0	0	0	0	0	0	24	0	0	0	0	[55]
GGIKCRTAYTSGATIACIGTCAT	ChenBR2	BR	676-698	1024	63.6-69.1	31	67	87	40	0	0	7	71	0	0	0	39	0	0	[52]

Table 4.2 continued.

						nifH [a] (%)			Specific groupings [b] (%)								Environ. [c] (%)			
Primer [d]	**Name**	**Tg.**[e]	**Pos.**[f]	**Deg.**[g]	**T_m (°C)**	**0**	**1**	**2**	***Pr***	***Cy***	**III**	**IA**	***Fr***	***Pb***	**Ep**	**IV**	**Soil**	**Mat**	**Sea**	**Ref.**[h]
GAAGACGATCCCGACCCCGA	FGPH750	Fr	759-778	1	66.8	0	1	1	0	0	0	0	25	0	0	0	0	0	0	[48]
AGCATGTCYTCSAGYTCNTCCA	nifHI	R	785-806	32	63.3-68.8	24	41	51	44	0	0	0	0	0	0	0	100	0	0	[53]
GGTCGGGACCTCATCCTCGA	FGPD913'	Fr	NA[i]	1	66.3	10	10	10	0	0	NA	NA	100	NA	NA	0	NA	NA	NA	[48]

[a] Data indicate primer binding to all *nifH* sequences in the database with 0, 1, and 2 mismatches allowed. In some cases highly degenerate primers bind to multiple positions in the sequence generating coverage values that exceed 100%.

[b] Data indicate primer binding to specific groupings in the *nifH* phylogeny. Abbreviations are as follows: *Alpha-, Beta-, and Gammaproteobacteria* (**Pr**); *Cyanobacteria* (**Cy**)**;** Cluster III (**III**); Cluster IA (**IA**); *Paenibacillus* (**Pb**); *Frankia* (**Fr**); *Epsilonproteobacteria* Containing Cluster (**Ep**); paralogous sequences in Cluster IV (**IV**).

[c] Primer coverage queried against sequences recovered from specific environments (Environ.) as described in methods. Environments include: soils (**Soil**), microbial mats (**Mat**), and pelagic marine samples (**Sea**).

[d] Sequences are given in the 5' to 3' direction, IUPAC characters are used, and I=Inosine.

[e] Abbreviations indicate the Target Group (Tg.) which the primer was intended to amplify as follows: β-*Rhizobia* (BR); *Frankia* (Fr); *Alphaproteobacteria* (AP); Symbiotic rhizobia (R); reamplification of Cluster I (RA); aerobic and microaerophilic diazotrophs (AM); *Cyanobacteria* (Cy); *Alpha- and Betaproteobacteria* (AB); *Gammaproteobacteria* (GP); alternative nitrogenase cluster (AN); designed to match multiple species of *Azospirillum*, *Burkholderia*, *Gluconoacetobacter*, *Azotobacter*, *Herbaspirillum* and *Azoarcus* (MS); species of the cyanobacterial genus *Cylindrospermopsis* (Cs); *Bradyrhizobium* sp. prevalent in truffles (TB).

[f] Position is relative to *A. vinelandii nifH* (Genbank ACCN# M20568).

[g] Degeneracy is given as the number of oligonucleotides that comprise the primer.

[h] References in which the primers are described.

[i] This binding position for this primer sequence lies beyond the stop codon of *Frankia* sp. (Genbank ACCN# M21132) and cannot be represented using the *A. vinelandii* numbering system.

NA: Data not available as described in Methods.

Table 4.3 Properties of universal primer pairs and their coverage for phylogenetic and environmental groupings in the *nifH* database.

Table 4.3

				Specific groupings [a] (%)								Environ. [b] (%)		
Primer set	**Pos.[c]**	**Len.[d]**	***nifH*[e]**	***Pr***	***Cy***	**III**	**IA**	***Fr***	***Pb***	**Ep**	**IV**	**Soil**	**Mat**	**Sea**
Nh21F/Cy55Nh428R	19-404	386	67	71	98	45	74	62	73	25	13	93	100	89
Ueda19F/407R	19-407	389	86	86	100	82	100	100	80	75	48	0	100	100
NH21F/nifH1	19-476	458	91	90	100	85	100	100	100	82	1	NA	100	100
nifH19F/nifH-univ463R	19-482	464	88	86	96	76	100	100	100	100	1	NA	100	100
Ueda19F/nifH-univ463r	19-482	464	87	86	96	76	100	100	100	82	1	NA	100	100
19F/nifH3	19-494	476	92	87	96	100	100	100	100	82	32	NA	100	100
Nh21F/nifH3	19-494	476	92	87	96	100	100	100	100	82	32	NA	100	100
nifH3/nifH4	22-494	473	49	68	96	0	0	100	100	27	14	NA	100	78
Primer-f/Primer-r	23-638	616	8	10	0	9	0	0	0	39	0	NA	0	0
FGPH19/FGPH273	25-279	255	3	3	0	1	0	71	0	0	0	0	0	0
PicenoF44/PicenoR436	25-453	429	33	52	2	5	6	85	0	2	0	12	0	0
F1/R6	25-473	449	85	88	100	61	94	92	100	93	6	94	100	100
FGPH19/PolR	25-476	452	6	6	0	0	7	77	0	0	0	0	0	0
F1/nifH3r	25-491	467	51	67	8	42	87	0	75	39	3	25	100	0
MehtaF/MehtaR	28-416	389	56	44	81	73	44	88	75	43	55	52	100	95
IGK/FGPH273'	31-279	249	9	12	0	1	7	35	0	0	1	8	0	0
IGK/DVV	31-413	383	83	84	85	86	68	87	89	88	70	78	100	85
IGK/VCG	31-419	389	86	86	84	91	90	87	78	96	37	74	100	100
nifHF/nifHRb	31-437	407	0	0	0	0	0	0	0	0	0	0	0	0
nifHF/nifHRc	31-437	407	3	4	0	2	0	0	0	0	0	1	0	0
IGK/primer-4=AQE	31-453	423	30	39	6	7	3	81	13	0	0	19	0	0
IGK/PolR	31-476	446	32	32	22	31	68	24	25	44	0	32	100	0
nifHF-Rösch/nifHR	31-491	461	26	17	34	10	42	66	50	33	0	56	100	0
IGK/YAA=nifH3	31-494	464	93	90	97	96	100	100	100	100	49	100	100	100
RL28/RL25	37-554	518	5	7	0	0	0	0	0	0	0	0	0	0
KAD3/DVV	106-413	308	66	84	83	14	81	96	64	77	1	54	100	25
KAD3/VCG	106-419	314	70	84	84	15	69	96	64	83	1	62	100	30
469/R6	112-473	362	82	79	92	69	76	98	53	96	9	61	100	80
469/nifH1	112-476	365	81	76	92	78	69	98	53	77	8	57	100	71
469/470	112-479	368	83	78	91	72	78	98	53	95	8	63	100	79
nifHFor/470	112-479	368	82	78	91	72	78	98	53	95	8	62	100	79
nif112/nifH-univ463R	112-482	371	39	28	64	53	48	31	47	73	4	46	60	64
nifB/nifHRev	112-482	371	17	8	63	21	14	20	27	13	0	12	0	64
PolF/primer-4=AQE	115-453	339	18	26	1	13	6	40	0	0	0	24	4	2
F2/R6	115-473	359	95	95	98	84	98	98	103	97	13	92	99	91
nifH2/R6	115-473	359	94	94	98	83	97	98	103	95	13	90	99	88
nifH1/nifH2	115-476	362	92	91	96	88	94	98	104	77	11	86	99	81
PolF/PolR	115-476	362	25	30	2	11	32	59	21	3	0	51	0	0
Kadino/Emino	115-482	368	83	87	51	67	79	98	87	97	7	84	100	76
Kadino/nifH-univ-463R	115-482	368	82	86	51	65	79	98	87	96	7	84	100	72
nifH11/nifH22	118-476	359	12	15	10	7	17	1	47	6	1	8	17	22
nifH-2f/nifH-3r	277-491	215	45	63	6	31	66	3	30	33	0	74	100	0

Table 4.3 continued.

				Specific groupings [a] (%)								Environ. [b] (%)		
Primer set	**Pos.[c]**	**Len.[d]**	***nifH*[e]**	***Pr***	***Cy***	**III**	**IA**	***Fr***	***Pb***	**Ep**	**IV**	**Soil**	**Mat**	**Sea**
Kadino/nifH-univ-463R	115-482	368	82	86	51	65	79	98	87	96	7	84	100	72
nifH11/nifH22	118-476	359	12	15	10	7	17	1	47	6	1	8	17	22
nifH-2f/nifH-3r	277-491	215	45	63	6	31	66	3	30	33	0	74	100	0

[a] Data indicate primer binding to specific groupings in the *nifH* phylogeny. Abbreviations are as follows: *Alpha, Beta, and Gamma Proteobacteria* (**Pr**); *Cyanobacteria* (**Cy**)**;** Cluster III (**III**); Cluster IA (**IA**); *Paenibacillus* (**Pb**); *Frankia* (**Fr**); *Epsilon Proteobacteria* Containing Cluster (**Ep**); paralogous sequences in Cluster IV (**IV**). In some cases highly degenerate primers bind to multiple positions in the sequence generating coverage values that exceed 100%.

[b] Primer coverage queried against sequences recovered from specific environments (Environ.) as described in methods. Environments include: soils (**Soil**), microbial mats (**Mat**), and pelagic marine samples (**Sea**).

[c] Position of amplicon in *nifH* is relative to *A. vinelandii nifH* (Genbank ACCN# M20568).

[d] Length expected for PCR amplicon.

[e] Data indicate primer binding with 0 mismatches to all *nifH* sequences in the database.

NA Data not available as nucleotide information is not available for the target group in the region of primer binding.

Table 4.4 Properties of group-specific primer pairs and their coverage for phylogenetic and environmental groups.

				Specific groupings [a] (%)								Environ. [b] (%)		
Primer set	**Pos.[c]**	**Len.[d]**	***nifH*[e]**	***Pr***	***Cy***	**III**	**IA**	***Fr***	***Pb***	**Ep**	**IV**	**Soil**	**Mat**	**Sea**
ChenBR1/ChenBR2	18-698	681	19	35	0	0	7	75	0	0	0	NA	0	0
nifH-a1-forA/nifH-a1-rev	19-482	464	5	12	0	0	0	0	0	0	0	NA	0	0
nifH-f1-forA/nifH-f1-rev	19-482	464	3	0	0	0	0	92	0	0	0	NA	0	0
nifHF/nifHI	25-806	782	15	33	0	0	0	0	0	0	0	100	0	0
primer-3/primer-4=AQE	37-453	417	0	0	0	0	0	0	0	0	0	0	0	0
nifHf1/nifH-269	43-341	299	1	0	0	0	0	19	0	0	0	0	0	0
nifHf1/nifHr	43-690	648	0	0	0	0	0	0	0	0	0	0	0	0
Olsen1/Olsen2	106-494	389	0	0	0	0	0	0	0	0	0	0	0	0
cyanoF/cyanoR	108-482	375	0	0	0	0	0	0	0	0	0	0	0	0
nifH-a1-forB/nifH-a1-rev	112-482	371	5	11	0	0	0	0	0	0	0	2	0	0
nifH-a2-forB/nifH-a2-rev	112-482	371	0	0	0	0	0	0	0	0	0	0	0	0
nifH-b1-forB/nifH-b1-rev	112-482	371	1	2	0	0	6	0	0	0	0	5	0	0
nifH-c1-forB/nifH-c1-rev	112-482	371	1	1	0	3	1	0	0	4	0	4	0	0
nifH-f1-forB/nifH-f1-rev	112-482	371	20	4	0	2	7	87	0	0	0	6	0	0
nifH-g1-forB/nifH-g1-rev	112-482	371	1	1	0	0	0	0	0	1	0	2	0	0
nifHFor/NifHRev	136-452	317	5	9	0	0	1	0	0	0	0	3	0	0
cylnif-F/cylnif-R	156-375	220	0	0	0	0	0	0	0	0	0	0	0	0
NIFH2F/NIFH5R	231-326	96	0	1	0	0	0	0	0	0	0	1	0	0
FGPH750/FGPD913'	759-[f]	116	0	0	0	NA	NA	0	NA	NA	0	NA	NA	NA

[a] Data indicate primer binding to specific groupings in the *nifH* phylogeny. Abbreviations are as follows: *Alpha, Beta, and Gamma Proteobacteria* (**Pr**); *Cyanobacteria* (**Cy**)**;** Cluster III (**III**); Cluster IA (**IA**); *Paenibacillus* (**Pb**); *Frankia* (**Fr**); *Epsilon Proteobacteria* Containing Cluster (**Ep**); paralogous sequences in Cluster IV (**IV**). In some cases highly degenerate primers bind to multiple positions in the sequence generating coverage values that exceed 100%.

[b] Primer coverage queried against sequences recovered from specific environments (Environ.) as described in methods. Environments include: soils (**Soil**), microbial mats (**Mat**), and pelagic marine samples (**Sea**).

[c] Position of amplicon in *nifH* is relative to *A. vinelandii nifH* (Genbank ACCN# M20568).

[d] Length expected for PCR amplicon.

[e] Data indicate primer binding with 0 mismatches to all *nifH* sequences in the database.

[f] This binding position for the reverse primer sequence lies beyond the stop codon of *Frankia* sp. (Genbank ACCN# M21132) and cannot be represented using the *A. vinelandii* numbering system.

NA Data not available as nucleotide information is not available for the target group in the region of primer binding.

set ChenBR1/ChenBR2 is designed to target β-*Rhizobia* but also hits 35% of the sequences within the *Alpha-, Beta-, and Gammaproteobacteria* and 75% of *Frankia* sequences. The *Frankia*-specific primer sets nifH-f1-forA/nifH-f1-rev and nifH-f1-forB/nifH-f1-rev hit 92% and 87% of *Frankia* respectively.

Primer sets with high *in silico* coverage were used for empirical tests. When tested with DNA from soil, the primer combinations nifH2/R6, nH21f/nifH, nifH1/nifH2, Ueda19f/univ463r, and nifH3/nH21f all produced PCR products of indiscriminate size producing smeared bands in gel electrophoresis and also produced an amplified product from *E. coli* indicating a lack of specificity for *nifH* under the amplification conditions tested (Table 4.5). The primer combinations F2/R6, IGK3/DVV, and Ueda 19F/388R produced a band of the expected size for a diverse range of genomic and soil DNA templates (Table 4.5), though Ueda 19F/388R was observed to produce an amplified product from *E. coli* indicating a lack of specificity for *nifH* under the amplification conditions tested. Overall, the primer pair IGK3/DVV produced the best performance in our empirical analysis, producing PCR products of the expected size from all nitrogen-fixing strains and soil DNA samples tested, while not generating PCR product from the negative controls or producing non-specific PCR products (Table 4.5).

Discussion

We report a comprehensive evaluation of *nifH* PCR primers. Our analysis of

Table 4.5 Empirical results of PCR using different *nifH* primer sets with DNA from isolates and soils.[a]

	AT (°C)[b]	Dv	Gu	Av	Fs	Ml	Kp	Xa	Rs	Rl	Pn	Ec	AS	LS	NT
F2/R6	51	-	+	+	-	+	+	-	ns	-	-	-	+	+	-
IGK3/DVV	58	+	+	+	+	+	+	+	+	+	+	-	+	+	-
Ueda19F/388R	51	+	+	+	-	+	+	+	+	ns	+	ns	+	+	-
nifH2/R6	44	ns	+	+	+	+	+	ns	-	+	-	ns	s	s	-
nH21f/nifH1	46	ns	ns	+	-							ns	s	s	-
nifH1/nifH2	46	ns	+	+	-							ns	s	s	-
Ueda19f/univ463r	46	+	ns	+	-	+	+	+	-	+	+	ns	s	s	-
nifH3/nH21f	41	ns	ns	+	-							ns	s	s	-

[a] DNA samples and their phylogenetic affiliation in the *nifH* phylogeny are: *Desulfovibrio vulgaris* Hildenborough (**Dv**), cluster III; *Geobacter uraniireducens* Rf4 (**Gu**), subcluster IA; A*zotobacter vinelandii* DJ (**Av**), *Alpha-, Beta- and Gamma-Proteobacteria; Frankia* sp. CcI3 (**Fs**), *Frankia*; *Mastigocladus laminosus* UTEX LB 1931 (**Ml**), *Cyanobacteria*; *Klebsiella pneumoniae* 342 (**Kp**), *Alpha-, Beta- and Gammaproteobacteria*; *Xanthobacter autotrophicus* Py2 (**Xa**), *Alpha-, Beta- and Gammaproteobacteria*; *Rhodobacter sphaeroides* 2.4.1 (**Rs**), *Alpha-, Beta- and Gammaproteobacteria*; *Rhizobium leguminosarium* bv. trifolii (**Rl**), *Alpha-, Beta- and Gammaproteobacteria; Polaromonas naphthalenivorans* CJ2 (**Pn**), *Alpha-, Beta- and Gammaproteobacteria*; *Eschericia coli* (**Ec**), genomic-DNA negative control; agricultural soil (**AS**); lawn soil (**LS**); No Template Control (**NT**). The symbols used are: product of the correct size (+), no product produced (-), non-specific amplification producing multiple bands or a single band of the wrong size (ns), a smeared band of indiscriminate size overlapping in size with the expected product (**s**). Blank cells indicate that the evaluation was not performed.

[b] Annealing Temperature (**AT**) used in PCR.

nifH primers reveals disparities in their sequence coverage. Variation in coverage is especially notable for primers designed to be universal, where 23 out of 51 target fewer than 50% of known *nifH* sequences and only 15 target more than 90% of sequences (Table 4.1). There could be several reasons for the disparity in primer coverage and specificity. Adequate primer design requires use of a sequence database representing the entire sequence diversity to be targeted by the primer. The number of sequences available in public databases has grown dramatically in recent years and earlier efforts at primer design were constrained in the past by the limited number and diversity of *nifH* sequences available. There is also a reasonable tendency to seek minimally degenerate primers due to undesirable effects that high levels of primer degeneracy can have on PCR performance. Decisions to lower degeneracy, however, could come at the cost of adequate coverage of target sequences.

Our efforts to evaluate universal *nifH* primers expand upon previous work to design universal primers for this gene. Marusina *et al.* designed *nifH* primers based upon a diverse set of *nifH* sequences and tested the resulting primers against DNA from cultivated strains [18]. The F2/R6 primer set they designed was one of the best performing in our comparison (Tables 4.3 and 4.5). Fedorov *et al.* later reexamined some of the primers of Marusina *et al.* because they found that primer R6 contained mismatches to certain methylobacterial *nifH* sequences, and they sought to design primers that included this group [19]. The coverage of their new primer, nifH-3r, however, is considerably lower than that of the original R6 primer matching 48% and

96% of *nifH* sequences respectively (Table 4.1). Poly *et al.* also designed a universal primer set, PolF/PolR, and showed that it amplified 19 of 19 test strains and worked well in soils [20]. However, the test strains they used consisted of *Alpha-, Beta-, and Gammaproteobacteria*, *Firmicutes* and *Actinobacteria* and did not include cluster IA, *Cyanobacteria*, or cluster III sequences. We found that the PolF/PolR primer set only encompassed 25% of *nifH* diversity in our database (Table 4.3).

By mapping the 51 universal primers to their complementary binding positions along the *A. vinelandii nifH* gene (Figure 4.1), it is evident that the majority of the primers correspond to conserved regions of the *nifH* gene that encode essential functions like the P-loop, Switch I, and Switch II (Figure 4.1; [22]). Sequence coverage is high in regions of universal primer binding (Figure 4.1), and the shape of the coverage profile suggests that primer sequences have not been trimmed from a large number of sequences. If this is indeed the case, then there could be some bias in our results since the sequence fidelity between primer and target can vary as a function of the specificity of PCR conditions. If primer sequences have replaced existing *nifH* polymorphism in database sequences, then the net result would be a bias towards overestimating primer coverage. This is a common problem in public sequence databases and illustrates the need for depositors to remove primer sequences prior to sequence deposition.

Some of the primer sequences we evaluated have unusually low coverage perhaps indicating that the published sequences contain errors, a phenomenon which is

not that uncommon as it has been noted in another review of primer sequences [23]. In particular, there appear to be errors in the sequences published for the primers YAA-poly, nifHRb, and röschF-1b [20, 24, 25]. In the case of primer YAA-poly it appears that the first part of the primer name "YAA" was appended to the 5' end of the primer sequence in [20] because the original YAA primer sequence does not have these nucleotides [26]. The coverage values for the original YAA primer (the one without the 5' YAA nucleotides) are actually those of the primer nifH3 (Table 2). For primers nifHRb and röschF-1b there appear to be single base pair errors in the primer sequences. If a single base pair mismatch is allowed for these primers it causes coverage to increase substantially (Table 4.1, Table 4.2). The primer röschF-1b [25] differs from the primer nifHF-Rösch [24] in that a G rather than a C is present at the 13th nucleotide from the 5' terminus. In addition, the primer AMR-R, though reported as a *nifH* primer [27], does not match *nifH* and thus appears to be erroneous.

We evaluate primer coverage *in silico* but it is important to point out that universal *nifH* PCR primers have been used under a wide range of reaction conditions and variation in annealing temperatures and cycle parameters will have dramatic impacts on actual primer performance. Lowering of PCR annealing temperature, for example, lowers reaction specificity and may permit amplification of templates with mismatches in the primer binding region. Notably, for many primer sets either a nested, touchdown, or stepdown PCR approach was needed to achieve amplification of *nifH* genes from environmental samples (e.g. [28, 29]). In Tables 1-3 we indicate

primer coverage with up to two mismatches to provide an indication of the potential effects that reducing reaction stringency may have on primer performance. In addition, there are several other factors which could impact the specificity and coverage realized using PCR primers at the bench relative to predictions made using sequence databases. These factors include primer dimerization [30], hairpin formation [31], GC content [32], the location of mismatches [33], and the thermodynamics of primer binding to template [34]. For example, mismatches at the 3' end of a primer may have a greater impact on specificity than those at the 5' end [33] and some methods of primer design exploit this tendency in order to increase primer coverage [35]. Thus, the real test of primer performance comes at the bench. We performed empirical assessment of coverage for primers which we found targeted 90% or more of sequences in the *nifH* database. The primer combinations F2/R6, IGK3/DVV, and Ueda 19F/388R performed well with DNA from a diversity of phylogenetic groups and from soil, with IGK3/DVV performing best of all. In contrast, the primer sets Ueda19f/univ463r and nifH1/nifH2 (ie: the Zehr-McReynolds primers) had mediocre performance with soils, producing smeared bands indicative of non-specific amplification, and producing a PCR product from negative controls (Table 4.5). All other primer combinations tested had drawbacks such as poor or no soil amplification and amplification of negative controls (Table 4.5).

There are several limitations to our approach which must be considered. First, only a few full-length *nifH* sequences are currently available and this lowers the

sequence diversity represented along the termini of the *nifH* gene (Figure 4.1). Hence, evaluation of primers that bind near the beginning or end of the alignment must be interpreted with care, especially for phylogenetic groups that are underrepresented in sequence databases. Likewise, *nifH* diversity remains poorly characterized in some and thus estimates primer performance in specific environments must also be interpreted with care when the number of sequences from those environments are small. As the number of sequenced genomes increases, full length *nifH* sequences from more diverse nitrogen fixers will become available aiding future efforts at primer design and analysis. Secondly, we have made no effort to assess coverage for nested and semi-nested reactions, which are common approaches. Nested amplification strategies, when coupled with low stringency reaction conditions, can allow investigators to amplify a wider diversity of templates than would be predicted through *in silico* analysis. Logically, however, *in silico* results from nested designs would always produce a reduction in coverage relative to a single primer set design.

Some of the universal *nifH* primers amplify paralogous genes not involved in nitrogen-fixation, for example cluster IV genes (Table 4.1). The *nifH* gene shares conserved regions with genes of cluster IV and cluster V which is involved in bacteriochlorophyll synthesis [13, 22]. We find that a substantial number of *nifH* universal primers will amplify cluster IV sequences (Table 4.1). It would therefore be wise for researchers interested in assessing the diversity and phylogeny of nitrogen-fixation genes from the environment to screen their sequences for the presence of

cluster IV and cluster V genes prior to OTU clustering.

Our work outlines a comprehensive approach to primer evaluation. Molecular-based studies are dependent on the effectiveness of the primer sets used to generate the sequence data which serves as our window to the microbial world. These results show that many supposedly universal primer sets miss significant portions of known *nifH* diversity. Several of the primers that performed well *in silico* were tested empirically against genomic DNA from a phylogenetically diverse set of strains. The primers that performed well both *in silico* and empirically should have the greatest utility in further studies of the *nifH* gene diversity in environmental samples.

Materials and Methods

Primer coverage analyses were performed using an updated version of our previously described *nifH* database [9]. The current version of the database contains 23,847 sequences, representing all *nifH* sequences available in Genbank as of July 14, 2010. The database was constructed using the ARB software package [36] as described in [9]. Alignment positions are numbered relative to the *Azotobacter vinelandii* gene sequence (Genbank ACCN# M20568). The environmental origins of sequences (Tables 1-5) were determined by keyword searches of the sequence records in the *nifH* database using ARB as described in [9]. The phylogenetic trees and sequence configurations for the environmental groups may be examined as part of the ARB *nifH* database used for this work which is available at http://www.css.cornell.edu/faculty/buckley/nifH_database_2010_07_14.arb.

We visualized the nucleotide representation of *nifH* sequence fragments within our *nifH* database relative to the *A. vinelandii nifH* sequence (Figure 4.1) by first exporting in FASTA format all *nifH* sequences from the ARB database using the *A. vinlandii nifH* sequence as a filter so that only positions in the alignment where *A. vinelandii nifH* had a nucleotide were exported. The FASTA file was then opened in BioEdit [37] where we could calculate a positional nucleotide numerical summary, and the total number of sequences containing sequence information was then plotted for each position in the alignment (Figure 4.1).

Primer coverage calculations were performed using the EMBOSS programs fuzznuc, dreg, and primersearch [38] to analyze sequence alignment data exported in FASTA format from our *nifH* database. The program fuzznuc calculates the number of sequences in a given alignment hit by a given primer. Mismatches, or fuzzy searches, are allowed by the program and were performed with the *nifH* evaluations (Table 4.1-2). The program primersearch was used for the evaluation of primer pairs (Tables 4 and 5). The program dreg was used to determine the number of records in an alignment that contained sequence data in the alignment region targeted by each primer or primer pair (Tables 1-5). However, because dreg eliminates the gap characters from the FASTA alignment file from ARB, the flanking gap characters were converted to the IUPAC character S, which is preserved by dreg, and the intervening gap characters were subsequently converted to the IUPAC character N. This allowed the original column positions from the ARB alignment to be maintained and reported

as output from dreg. To calculate primer and primer pair coverage, the number of hits obtained from fuzznuc or primersearch were divided by the total number of sequences with nucleotide representation in the target region(s) as indicated by dreg.

Unix bash shell scripts were employed to increase the throughput of the *in silico* primer evaluations by automating the input of multiple primer sequences and other evaluation parameters into the EMBOSS programs. The scripts were also used to parse the output files and organize the data into tables.

Primer annealing temperatures were calculated with SciTools Oligoanalyzer version 3.1 which calculates oligonucleotide melting temperatures based on nearest neighbor thermodynamics [39]. Oligoanalyzer can account for Inosine but not for P or K bases and thus melting temperatures were not calculated for PicenoF44 and PicenoR436 (Table 4.1). The parameters used for the calculations were 0.25 μM oligonucleotides, 50 mM Na^{+}, 1.5 mM Mg^{++}, and 0 mM dNTPs.

Genomic DNA was extracted from cultures of the bacterial strains listed in Table 4.5 according to a standard enzymatic, phenol-chloroform extraction protocol [40]. DNA concentration was determined with a Nanodrop model 1000 (Thermo Fischer Scientific, Wilmington, DE), and DNA was diluted to 1 ng μl^{-1} prior to PCR. Soil DNA was obtained from a long-term agricultural site at the William H. Miner institute, Chazy, NY described previously [41]. The agricultural soil sample comes from a tilled site used to grow corn for more than 30 years while the lawn soil sample is from a non-cultivated control site that is adjacent to the agricultural site and

contains a mixed community of perennial grasses (Table 4.5). Soil samples were obtained by coring at 0-5 cm depth. Soil samples were sieved to 2 mm, frozen in the field using liquid nitrogen, and stored at -80° C. DNA was extracted from soils using the PowerSoil DNA Isolation Kit (MoBio, Carlsbad, CA).

Primers were synthesized and desalted by Integrated DNA technologies. All PCR reaction volumes were 50 µL with the following final reagent concentrations: 1X PCR Gold Buffer (ABI, Foster City, CA), 2.5 mM $MgCl_2$ solution (ABI, Foster City, CA), 0.05% BSA (NEB, Ipswich, MA), 0.2 mM dNTPs, 1 µM each primer, 2.5 U Amplitaq Gold DNA polymerase (ABI, Foster City, CA). As template, 1 ng of genomic DNA was added, or 1 µl of soil DNA extract. To visualize the PCR products, 10 µL of the reactions were loaded onto a 50 ml, 1% agarose gel with 1 µL of SYBR Safe dye (Molecular Probes, Eugene, OR). 5 µl of Hyperladder I (Bioline, Taunton, MA) was loaded onto each gel as a molecular weight marker. Gels ran for 45 minutes at 100 volts and 500 miliamps and were then visualized and photographed.

References

1. Zehr JP, Jenkins BD, Short SM, Steward GF (2003) Nitrogenase gene diversity and microbial community structure: a cross-system comparison. Environ Microbiol 5: 539-554.

2. Rubio LM, Ludden PW (2002) The gene products of the *nif* regulon. In: Leigh GJ, editor. Nitrogen fixation at the millennium. Elsevier Science B.V. pp. 101-136.

3. Man-Aharonovich D, Kress N, Bar Zeev E, Berman-Frank I, Beja O (2007) Molecular ecology of *nifH* genes and transcripts in the eastern Mediterranean Sea. Environ Microbiol 9: 2363.

4. Roesch C, Bothe H (2009) Diversity of total, nitrogen-fixing and denitrifying bacteria in an acid forest soil. Eur J Soil Sci 60: 883-894.

5. Mehta MP, Butterfield DA, Baross JA (2003) Phylogenetic diversity of nitrogenase (*nifH*) genes in deep-sea and hydrothermal vent environments of the Juan de Fuca Ridge. Appl Environ Microbiol 69: 960-970.

6. Héry M, Philippot L, Mériaux E, Poly F, Le Roux X *et al.* (2005) Nickel mine spoils revegetation attempts: effect of pioneer plants on two functional bacterial communities involved in the N-cycle. Environ Microbiol 7: 486-498.

7. Yamada A, Inoue T, Noda S, Hongoh Y, Ohkuma M (2007) Evolutionary trend of phylogenetic diversity of nitrogen fixation genes in the gut community of wood-feeding termites. Mol Ecol 16: 3768-3777.

8. Roesch L, Camargo F, Bento F, Triplett E (2008) Biodiversity of diazotrophic bacteria within the soil, root and stem of field-grown maize. Plant Soil 302: 91-104.

9. Gaby JC, Buckley DH (2011) A global census of nitrogenase diversity. Environ Microbiol 13: 1790-1799.

10. Reed SC, Townsend AR, Cleveland CC, Nemergut DR (2010) Microbial community shifts influence patterns in tropical forest nitrogen fixation. Oecologia 164: 521-531.

11. Hsu S, Buckley DH (2009) Evidence for the functional significance of diazotroph community structure in soil. ISME J 3: 124-136.

12. Chien Y, Zinder SH (1994) Cloning, DNA-sequencing, and characterization of a *nifD*-homologous gene from the archaeon *Methanosarcina barkeri* 227 which resembles *nifD1* from the eubacterium *Clostridium pasteurianum*. J Bacteriol 176: 6590-6598.

13. Raymond J, Siefert JL, Staples CR, Blankenship RE (2004) The natural history of nitrogen fixation. Mol Biol Evol 21: 541-554.

14. Young JPW (2005) The phylogeny and evolution of nitrogenases. In: Palacios R, Newton WE, editors. Genomes and genomics of nitrogen-fixing organisms. Springer. pp. 221-241.

15. Young JPW (1992) Phylogenetic classification of nitrogen-fixing organisms. In: Stacey G, Burris RH, Evans HJ, editors. Biological nitrogen fixation. Chapman and Hall. pp. 43-86.

16. Zehr JP, McReynolds LA (1989) Use of degenerate oligonucleotides for amplification of the *nifH* gene from the marine cyanobacterium *Trichodesmium thiebautii*. Appl Environ Microbiol 55: 2522-2526.

17. Kirshtein JD, Paerl HW, Zehr J (1991) Amplification, cloning, and sequencing of a *nifH* segment from aquatic microorganisms and natural communities. Appl Environ Microbiol 57: 2645-2650.

18. Marusina AI, Boulygina ES, Kuznetsov BB, Tourova TP, Kravchenko IK *et al.* (2001) A system of oligonucleotide primers for the amplification of *nifH* genes of different taxonomic groups of prokaryotes. Mikrobiologiya 70: 86-91.

19. Fedorov DN, Ivanova EG, Doronina NV, Trotsenko IA (2008) A new system of degenerate-oligonucleotide primers for detection and amplification of *nifHD* genes. Mikrobiologiia 77: 286-288.

20. Poly F, Monrozier LJ, Bally R (2001) Improvement in the RFLP procedure for studying the diversity of *nifH* genes in communities of nitrogen fixers in soil. Res Microbiol 152: 95-103.

21. Bürgmann H, Widmer F, Von Sigler W, Zeyer J (2004) New molecular screening tools for analysis of free-living diazotrophs in soil. Appl Environ Microbiol 70: 240-247.

22. Schlessman JL, Woo D, Joshua-Tor L, Howard JB, Rees DC (1998) Conformational variability in structures of the nitrogenase iron proteins from *Azotobacter vinelandii* and *Clostridium pasteurianum*. J Mol Biol 280: 669-685.

23. Arahal DR, Llop P, Alons MP, Lopez MM (2004) *In silico* evaluation of molecular probes for detection and identification of *Ralstonia solanacearum* and *Clavibacter michiganensis* subsp sepedonicus. Syst Appl Microbiol 27: 581-591.

24. Rösch C, Bothe H (2005) Improved assessment of denitrifying, N_2-fixing, and total-community bacteria by terminal restriction fragment length polymorphism analysis using multiple restriction enzymes. Appl Environ Microbiol 71: 2026-2035.

25. Ogilvie LA, Hirsch PR, Johnston AWB (2008) Bacterial diversity of the Broadbalk 'classical' winter wheat experiment in relation to long-term fertilizer inputs. Microb Ecol 56: 525-537.

26. Ohkuma M, Noda S, Usami R, Horikoshi K, Kudo T (1996) Diversity of nitrogen fixation genes in the symbiotic intestinal microflora of the termite *Reticulitermes speratus*. Appl Environ Microbiol 62: 2747-2752.

27. Rösch C, Mergel A, Bothe H (2002) Biodiversity of denitrifying and dinitrogen-fixing bacteria in an acid forest soil. Appl Environ Microbiol 68: 3818-3829.

28. Hewson I, Moisander PH, Morrison AE, Zehr JP (2007) Diazotrophic bacterioplankton in a coral reef lagoon: phylogeny, diel nitrogenase expression and response to phosphate enrichment. ISME J 1: 78-91.

29. Duc L, Noll M, Meier BE, Bürgmann H, Zeyer J (2009) High diversity of diazotrophs in the forefield of a receding alpine glacier. Microb Ecol 57: 179-190.

30. Brownie J, Shawcross S, Theaker J, Whitcombe D, Ferrie R *et al.* (1997) The elimination of primer-dimer accumulation in PCR. Nucleic Acids Res 25: 3235-3241.

31. Singh VK, Govindarajan R, Naik S, Kumar A (2000) The effect of hairpin structure on PCR amplification efficiency. Mol Biol Today 1: 67-69.

32. Benita Y, Oosting RS, Lok MC, Wise MJ, Humphery-Smith I (2003) Regionalized GC content of template DNA as a predictor of PCR success. Nucleic Acids Res 31: e99.

33. Bru D, Martin-Laurent F, Philippot L (2008) Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. Appl Environ Microbiol 74: 1660-1663.

34. Polz M, Cavanaugh C (1998) Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol 64: 3724-3730.

35. Rose TM, Henikoff JG, Henikoff S (2003) CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Res 31: 3763-3766.

36. Ludwig W, Strunk O, Westram R, Richter L, Meier H *et al.* (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32: 1363-1371.

37. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41: 95-98.

38. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276-277.

39. Owczarzy R, Tataurov AV, Wu Y, Manthey JA, McQuisten KA *et al.* (2008) IDT SciTools: a suite for analysis and design of nucleic acid oligomers. Nucleic Acids Res 36: W163-9.

40. Yeates C, Gillings MR, Davison AD, Altavilla N, Veal DA (1998) Methods for microbial DNA extraction from soil for PCR amplification. Biol Proced Online 1: 40-47.

41. Moebius BN, van Es HM, Schindelbeck RR, Idowu OJ, Clune DJ *et al.* (2007) Evaluation of laboratory-measured soil properties as indicators of soil physical quality. Soil Sci 172: 895-912.

42. Deslippe JR, Egger KN, Henry GHR (2005) Impacts of warming and fertilization on nitrogen-fixing microbial communities in the Canadian High Arctic. FEMS Microbiol Ecol 53: 41-50.

43. Widmer F, Shaffer BT, Porteous LA, Seidler RJ (1999) Analysis of *nifH* gene pool complexity in soil and litter at a Douglas Fir forest site in the Oregon Cascade Mountain Range. Appl Environ Microbiol 65: 374-380.

44. Ueda T, Suga Y, Yahiro N, Matsuguchi T (1995) Remarkable N_2-fixing bacterial diversity detected in rice roots by molecular evolutionary analysis of *nifH* gene sequences. J Bacteriol 177: 1414-1417.

45. Ando S, Goto M, Meunchang S, Thongra-ar P, Fujiwara T *et al.* (2005) Detection of *nifH* sequences in sugarcane (*Saccharum officinarum* L.) and pineapple (*Ananas comosus* [L.] Merr.). Soil Sci Plant Nutr 51: 303-308.

46. Zani S, Mellon MT, Collier JL, Zehr JP (2000) Expression of *nifH* genes in natural microbial assemblages in Lake George, New York, detected by reverse transcriptase PCR. Appl Environ Microbiol 66: 3119-3124.

47. Flores-Mireles AL, Winans SC, Holguin G (2007) Molecular characterization of diazotrophic and denitrifying bacteria associated with mangrove roots. Appl Environ Microbiol 73: 7308-7321.

48. Simonet P, Grosjean MC, Misra AK, Nazaret S, Cournoyer B *et al.* (1991) *Frankia* genus-specific characterization by polymerase chain reaction. Appl Environ Microbiol 57: 3278-3286.

49. Piceno Y, Noble P, Lovell C (1999) Spatial and temporal assessment of diazotroph assemblage composition in vegetated salt marsh sediments using denaturing gradient gel electrophoresis analysis. Microb Ecol 38: 157-167.

50. Bagwell CE, Piceno YM, Ashburne-Lucas A, Lovell CR (1998) Physiological diversity of the rhizosphere diazotroph assemblages of selected salt marsh grasses. Appl Environ Microbiol 64: 4276-4282.

51. Yeager CM, Kornosky JL, Housman DC, Grote EE, Belnap J *et al.* (2004) Diazotrophic community structure and function in two successional stages of biological soil crusts from the Colorado Plateau and Chihuahuan Desert. Appl Environ Microbiol 70: 973-983.

52. Chen W, James EK, Chou J, Sheu S, Yang S *et al.* (2005) β-Rhizobia from *Mimosa pigra*, a newly discovered invasive plant in Taiwan. New Phytol 168: 661-675.

53. Laguerre G, Nour SM, Macheret V, Sanjuan J, Drouin P *et al.* (2001) Classification of rhizobia based on *nodC* and *nifH* gene analysis reveals a close phylogenetic relationship among *Phaseolus vulgaris* symbionts. Microbiology 147: 981-993.

54. Lovell CR, Piceno YM, Quattro JM, Bagwell CE (2000) Molecular analysis of diazotroph diversity in the rhizosphere of the smooth cordgrass, *Spartina alterniflora*. Appl Environ Microbiol 66: 3814-3822.

55. Mirza BS, Welsh A, Rasul G, Rieder JP, Paschke MW *et al.* (2009) Variation in *Frankia* populations of the *Elaeagnus* host infection group in nodules of six host plant species after inoculation with soil. Microb Ecol 58: 384-393.

56. Olson J, Steppe T, Litaker R, Paerl H (1998) N_2-fixing microbial consortia associated with the ice cover of Lake Bonney, Antarctica. Microb Ecol 36:

231-238.

57. Soares RA, Roesch LFW, Zanatta G, de Oliveira Camargo FA, Passaglia LMP (2006) Occurrence and distribution of nitrogen fixing bacterial community associated with oat (*Avena sativa*) assessed by molecular and microbiological techniques. Applied Soil Ecology 33: 221-234.

58. Dyble J, Paerl HW, Neilan BA (2002) Genetic characterization of *Cylindrospermopsis raciborskii* (cyanobacteria) isolates from diverse geographic origins based on *nifH* and *cpcBA*-IGS nucleotide sequence analysis. Appl Environ Microbiol 68: 2567-2571.

59. Barbieri E, Ceccaroli P, Saltarelli R, Guidi C, Potenza L *et al.* (2010) New evidence for nitrogen fixation within the Italian white truffle *Tuber magnatum*. Fungal Biol 114: 936-942.

60. Mirza BS, Welsh A, Rieder JP, Paschke MW, Hahn D (2009) Diversity of frankiae in soils from five continents. Syst Appl Microbiol 32: 558-570.

CHAPTER 5

OPTIMIZATION OF *NIFH* PRIMERS FOR QUANTITATIVE PCR REVEALS THAT THE USE OF DEGENERATE PRIMERS CAN CAUSE DRAMATIC QUANTIFICATION BIAS

Introduction

Real-time quantitative PCR (qPCR) has emerged as a powerful technique to enumerate gene abundance in many fields of biology. In the field of microbial ecology, the technique permits study of the environmental distribution of ecologically important microbes (e.g. [1]). The qPCR approach has been used to quantify the abundance of specific phylogenetic groups by targeting 16S rRNA genes as well as functional groups by targeting essential genes within metabolic pathways. The qPCR approach is often used to relate the abundance of microbial functional groups to physical and chemical factors of the environment. Protein-encoding genes have much greater variability in nucleotide sequence than 16S rRNA genes, and thus PCR primers that target functional genes often need to be degenerate (e.g. [2, 3]). The use of degenerate primers in PCR can be problematic for a variety of reasons but the implications of degenerate primer use in qPCR applications has not been thoroughly investigated.

PCR can be subject to amplification bias in which certain templates are

preferentially amplified resulting in final product ratios that differ from initial template ratios [4]. While there has been much research assessing the bias of primers specific to SSU rRNA sequences [5-14] , there are fewer studies that examine the potential biases caused by using degenerate primers to amplify functional genes (e.g. [15-17]). A universal primer for a functional gene such as *nifH* may possess more than 100-fold degeneracy (e.g. [18-20]). In PCR reactions with degenerate primers, templates with G-C base pairs at the degenerate positions tend to be favored over those with A-T base pairs [5]. In a study on methanogens it was found that changing the annealing temperature of the PCR reaction had a significant impact on relative product ratio as measured by TRFLP when using degenerate primers for the *mcrA* gene though there was little effect when using minimally degenerate primers that target the 16S rRNA gene [21]. The degree to which primer degeneracy impacts PCR amplification bias and the implications of such bias for qPCR remain poorly characterized.

Amplification bias may pose a serious concern for qPCR applications because quantification results could be skewed simply due to differences in the phylogenetic composition of microbial communities. For example, when a community contains a greater proportion of ’favored’ templates (i.e. templates with an amplification efficiency greater than the controls used for the standard curve), preferential amplification of the favored templates in qPCR would yield an inflated number of gene copies. Likewise, changes in the proportion of ‘unfavored’ templates (i.e. templates with an amplification efficiency less than the controls used for the standard

curve) would yield underestimates of gene copy number. Thus, we hypothesize that qPCR amplification bias can cause copy number estimates to vary as a function both of gene copy number and change in community composition.

In this article we show that the use of degenerate primers in qPCR can cause dramatic bias in the quantification of gene copy number. We performed analyses using degenerate primers that target *nifH*. While there are at least 50 universal *nifH* primers [2] with varying degrees of degeneracy, only a few of these primer sets have been used for qPCR. We selected several of the *nifH* primers used most commonly in qPCR and performed a systematic assessment of the potential for qPCR bias. The approach we demonstrate can be used to validate the performance of degenerate primers in qPCR applications. Using this approach we show that most *nifH* primers that have been used previously in qPCR have a strong bias with respect to the phylogenetic composition of communities. We identified a *nifH* qPCR primer set that was unbiased. This primer set was then used to quantify *nifH* copy number in an agricultural experiment, revealing that tillage causes a significant increase in *nifH* gene copy number. We further provide evidence that this result may be associated with organic matter loss in these systems. The approach we describe for validating qPCR primers is relevant to any application of qPCR to quantify functional gene copy number where degenerate PCR primers are required.

Materials and Methods

Preparation of *nifH* gene standards

A set of 5 strains (*Frankia* sp. CcI3, *Klebsiella pneumonia* 342, *Geobacter uraniireducens* RF4, *Desulfovibrio vulgaris* Hildenborough, and *Mastigocladus laminosus* UTEX LB 1931) were selected to represent a wide spectrum of phylogenetic diversity in *nifH*. These strains encompass considerable sequence variation in the region of primer binding (Table 5.1). A genome sequence is available for each of these strains which allowed us to design strain-specific PCR primers (Table 5.1) to amplify the entire length of each *nifH* gene encompassing the start and stop codons. Genomic DNA from the 5 representatives was used as a template in PCR reactions to generate full-length *nifH* gene standards. PCR was performed in a 50 µl volume with 2.5 units of Amplitaq gold DNA polymerase (Applied Biosystems, Foster City, CA), 1 µl each genomic DNA template, 1X PCR Gold buffer (Applied Biosystems, Carlsbad, CA), 2.5 mM $MgCl_2$ (Applied Biosystems, Carlsbad, CA), 200 µM dNTPs (Promega, Madison, WI), and 200 nm of each primer. Cycling parameters consisted of a 95°C hot start for 10 minutes followed by 35 cycles of a 3-step PCR consisting of 95 °C for 30 s, the annealing temperature (Table 5.1) for 30 s, and a 72 °C extension step for 45 s. The PCR cycling concluded with a final 10 minute extension at 72 °C. The PCR-amplified *nifH* gene products were purified using a

Table 5.1: Primer sequences, their T_m, and the annealing temperature used for PCR-amplification of the *nifH* gene standards.

species name	strain name	forward primer (5' to 3')	T_m forward primer (°C)	reverse primer (5' to 3')	T_m reverse primer (°C)	Anneal.[b] (°C)
Desulfovibrio vulgaris	Hildenborough	CGC CAT GAG AAA GGT AGC CAT CTA	58.5	GGC CGC GCT ACG ACG CTT	64.1	61.0
Frankia species	CcI3	ATG CGC CAG ATC GCA TTC TAT	56.9	TGG TCG GGA CCT CAT CCT CGA	62.0	58.0
Mastigocladus laminosus[a]	LB 1931	ATG ACT GAG AAT ATC AGA CA	47.7	TTA TTT AGC AGA AGC TTC A	46.1	50.0
Geobacter uraniireducens	RF4	ATG AGA CAG ATC GCG ATT TA	51.1	GCC CCT AAA TCA CCA GAT TAT TTT	53.4	56.0
Klebsiella pneumoniae	342	ATG ACC ATG CGT CAA TGC GC	59.1	TCA GGC CGC GTT TTC TTC AG	58.2	59.5

[a] To alleviate weak amplification of the *Mastigocladus laminosus nifH* standard, the Mg concentration of the PCR reaction was increased to 3.5 mM, the genomic DNA template was increased to 5 µl, and the number of PCR cycles extended to 41.

[b] Annealing temperature used for the PCR reaction.

Wizard SV Gel and PCR Clean-up System kit (Promega, Madison, WI), and DNA concentrations were determined using a Quanti-iT PicoGreen dsDNA assay kit (Invitrogen, Eugene, Oregon).

The amplification of *nifH* gene standards from the genomic DNA samples resulted in amplification products of the expected size (~900 bp) for each of the standards. For additional verification, we sequenced the 5 *nifH* gene standards using the forward primer that was used to amplify the standards, and in all 5 cases the amplified products matched the sequence of the *nifH* gene from their derivative genome.

qPCR assessment of primer sets for template-specific bias

We tested the universal *nifH* primer sets PolF/PolR [22], Ueda19F/Ueda407R [23], nifH1/nifH2 [18], F2/R6 [20], and DVV/IGK3 [24] (Table 5.2) for preferential amplification of *nifH* templates by using the 5 *nifH* gene standards described above. These primers differ in their degree of degeneracy. The primer IGK3 compliments the greatest diversity of *nifH* sequence variants; this primer consists of a mix of 72 oligonucleotides targetting73,728 sequence variants. Its partner, the primer DVV consists of 8 oligonucleotides targetting 8,192 sequence variants. The PolF/PolR pair targets the fewest sequence variants at 24 and 8. The pair nifH1/nifH2 are the most degenerate consisting of 96 and 128 oligonucleotides, while the pair F2/R6 is the least

Table 5.2: Universal *nifH* primer sequences and the corresponding template from the *nifH* gene standards.

primer name (position[a])	PolF (115-134)	PolR (457-476)	Ueda19F (19-38)	Ueda407R (388-407)
primer sequence (5' to 3')	TGCGA**Y**CC**S**AA**R**GC**B**GACTC	AT**S**GCCATCAT**Y**TC**R**CCGGA	GC**IW**T**Y**TA**Y**GG**I**AA**R**GG**I**GG	AA**I**CC**R**CC**R**CA**I**AC**I**AC**R**TC
Kpn 342[b]	-----T--G--A--G-----	--C--------C--G-----	--TA-C--C--T--A--C--	--G--G--G--G--C--G--
Fsp CcI3[b]	-----C--C--G--C-----	--C--------C--G-----	--AT-C--T--C--G--T--	--C--G--G--G--G--G--
Gur RF4[b]	-----C--C--G--G-----	--G--------C--G-----	--GA-T--C--C--A--T--	--C--G--G--G--A--G--
Dvu Hildenborough[b]	-----C--C--G--C-----	--G--------C--G-----	--CA-C--C--C--G--C--	--G--G--G--C--C--G--
Mla UTEX LB 1931[b]	-----C--C--G--T-----	--C--------T--A-----	--TT-C--C--T--A--C--	--A--A--A--T--A--G--

primer name (position[a])	F2 (115-131)	R6 (457-473)	IGK3 (19-47)	DVV (388-413)
primer sequence (5' to 3')	TG**Y**GA**Y**CC**I**AA**I**GC**I**GA	GCCATCAT**Y**TC**I**CC**I**GA	GC**IW**THTA**Y**GG**I**AA**R**GG**I**GG**I**AT**H**GG**I**AA	AT**I**GC**R**AA**I**CC**I**CC**R**CA**I**AC**I**AC**R**TC
Kpn 342[b]	--C--T--G--A--G--	--------C--G--G--	--TA----C--T--A--C--T--C--T--	--G--G--G--G--G--G--C--G--
Fsp CcI3[b]	--C--C--C--G--C--	--------C--G--G--	--AT----T--C--G--T--T--C--C--	--C--G--C--G--G--G--G--G--
Gur RF4[b]	--C--C--C--G--G--	--------C--G--G--	--GA----C--C--A--T--C--C--C--	--G--G--C--G--G--G--A--G--
Dvu Hildenborough[b]	--C--C--C--G--C--	--------C--G--G--	--CA----C--C--G--C--C--C--C--	--G--G--G--G--G--C--C--G--
Mla UTEX LB 1931[b]	--C--C--C--G--T--	--------T--A--G--	--TT----C--T--A--C--T--C--T--	--A--G--A--A--A--T--A--G--

primer name (position[a])	nifH2 (115-131)	nifH1 (460-476)
primer sequence (5' to 3')	TG**Y**GA**Y**CC**N**AA**R**GC**N**GA	A**DN**GCCATCAT**Y**TC**N**CC
Kpn 342[b]	--C--T--G--A--G--	-TC--------C--G--
Fsp CcI3[b]	--C--C--C--G--C--	-TC--------C--G--
Gur RF4[b]	--C--C--C--G--G--	-TG--------C--G--
Dvu Hildenborough[b]	--C--C--C--G--C--	-TG--------C--G--
Mla UTEX LB 1931[b]	--C--C--C--G--T--	-TC--------T--A--

[a] Relative to the *nifH* sequence of *Azotobacter vinlandii* (Genbank ACCN# M20568)

[b] The following abbreviations represent species names: Kpn, *Klebsiella pneumoniae*; Fsp, *Frankia* species, Gur, *Geobacter uraniireducens*; Dvu, *Desulfovibrio vulgaris*; and Mla, *Mastigocladus laminosu*s. The - symbol represents conserved positions whose nucleotides match those of the primer sequence above.

volume of 25 µL and had final concentrations at 1X Power SYBR Green master mix (Applied Biosystems, Foster City, CA), 300 nM each primer, and 5 µl of the *nifH* templates at $5.81x10^4$ copies/µl. Primers were HPLC-purified. The 2-step cycling protocol used for qPCR consisted of an initial hot-start at 95 °C for 10 minutes, degenerate composed of 4 and 2 oligonucleotides. The qPCR reactions had a total followed by 40 cycles of 95 °C for 15 s and the annealing temperature for 60 s. Annealing temperatures ranging from 52 – 67 °C were assessed. The qPCR reactions were performed on a BioRad Cfx96 qPCR machine with a Biorad C1000 thermocycler. Ct values are reported as the mean of duplicate qPCR reactions.

Application of PolF/PolR primers to evaluate *nifH* gene abundance in soil

Soil samples were obtained from a continuous maize cultivation trial established in 1973 at the William H. Miner Institute for Agricultural Studies in Chazy, NY, USA. The trial consists of a 4 x 4 randomized complete block design that is a combination of till/no-till and corn stover biomass removal/retention (as described in [26]). There is an adjacent, non-cultivated area of mown grass that serves as a control. Each of 4 replicate blocks within a treatment was sampled by taking ten 2.5 cm diameter, 5 cm deep soil cores which were homogenized by sieving to 4mm to represent each replicate block. The soils were sampled on September 28, 2006 when they were flash frozen with liquid nitrogen within an hour of sampling, and then were

subsequently stored at -20 °C until DNA extraction on July 1, 2011.

DNA was extracted from 0.25 g of soil using a Powersoil DNA Isolation Kit (MoBio, Carlsbad, CA) according to the manufacturer's protocol except that we used 2 minutes of bead beating in a Mini BeadBeater-8 (Biospec Products, Bartlesville, OK) on the homogenize setting. We made dilution series with soil DNA samples in order to determine the dilution required to eliminate inhibitory effects from environmental contaminants which have been described elsewhere [27, 28]. We found that a 100-fold dilution eliminated inhibitory effects, and thus all soil DNA samples were diluted 100-fold prior to analysis. We used the PolF/PolR primer set with the Qiagen Quantifast SYBR Green mastermix because an assessment of different commercial master mixes (data not shown) revealed this one to have the highest sensitivity for our reaction conditions. All other reaction conditions were as reported above for qPCR. We ran three separate qPCR reactions using the soil DNA extract from each replicate block as template. No template controls and a dilution series of the *K. pneumoniae* 342 standard were included in every qPCR run.

The statistical software R (version 2.13.1, The R Foundation for Statistical Computing) was used to conduct both a one-way ANOVA and a two-way ANOVA with interaction.

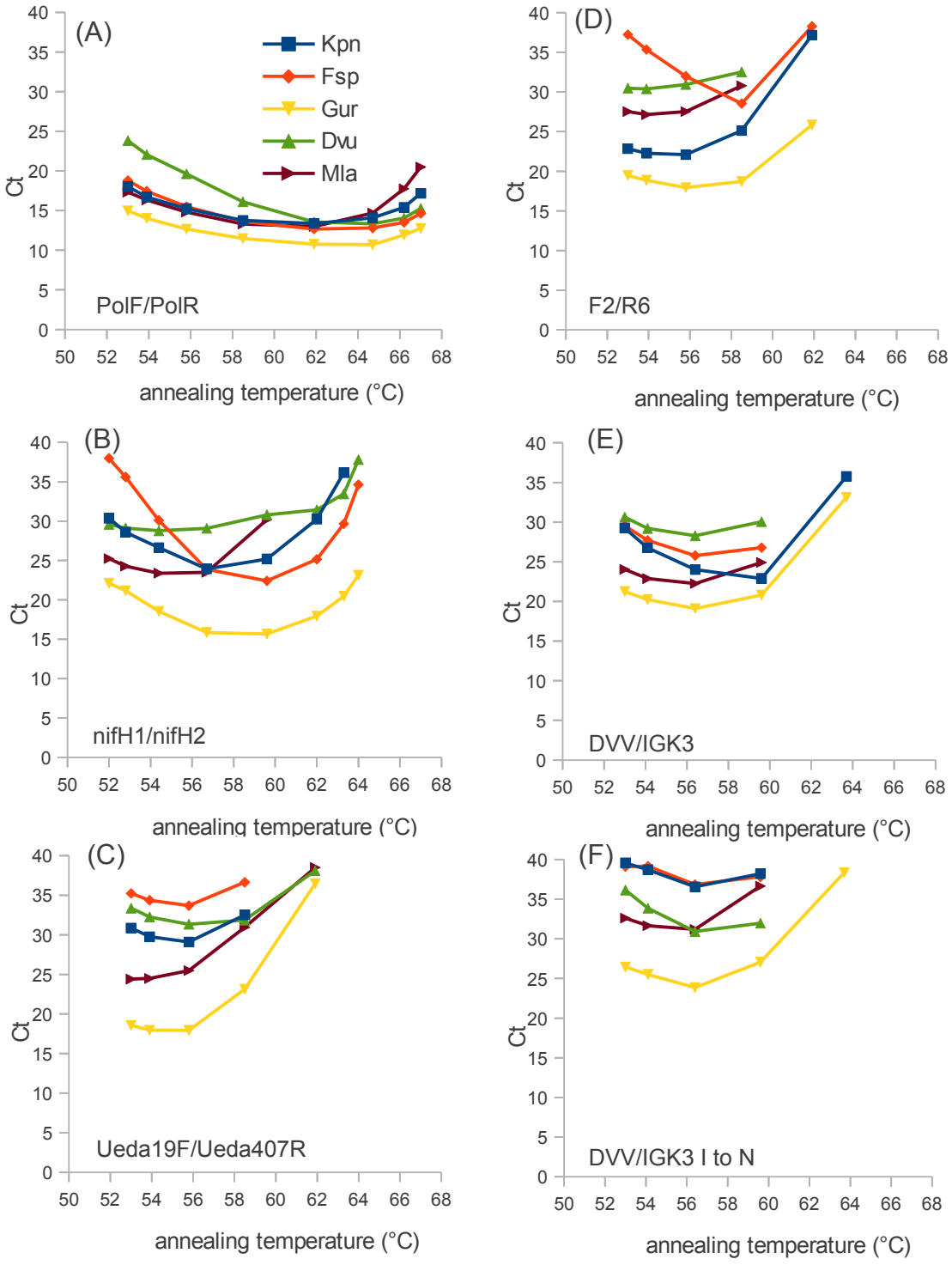
Results

assessment of template-specific qPCR bias

Amplification efficiency with degenerate primers varied dramatically across the 5 different templates tested (Figure 5.1). Samples that contained equal numbers of gene copies were observed to yield Ct values that differed by as much as 10 cycles, which would correspond to a 2^{10}, or 1024-fold, difference in the gene copy number estimate (Figure 5.1). The Ct values obtained for a particular primer set vary with both template identity and the annealing temperature.

Most universal *nifH* primer sets exhibited template-specific bias. In the case of those biased primer sets, the amplification efficiencies for each template differ resulting in Ct values that are vastly different for each template despite that there are equal copy numbers of each template. For instance, with the primer set nifH1/nifH2, at an annealing temperature of about 60 ºC, the *G. uraniireducens* template had the best amplification efficiency with a Ct of 16 whereas the template with the worst efficiency, *D. vulgaris* had a Ct of 31 despite the fact that the two were set to an equal copy number based upon PicoGreen assay quantification. The remaining templates had values that varied between these two extremes, with the *Frankia* species at a Ct of 22, *K. pneumoniae* at 25, and *M. laminosus* at 30. The Ct value for *G. uraniireducens* was numerically lower in all of the assessments indicating an amplification efficiency better than that of all the others. For other templates there was not a consistent order of appearance in the Ct profiles. For example, the *Frankia* species template had the

Figure 5.1: Evaluation of 6 universal *nifH* primer sets against 5 phylogenetically diverse *nifH* gene standards. Each reaction contained 2.9 x 10^5 *nifH* copies. The primer sets are labeled in the panels, and their sequences are given in Table 5.2, except for "DVV/IGK3 I to N" which is a modified version of DVV/IGK3 where the primer Inosines were changed to N's which are a mix the 4 normal nucleotides. The *nifH* gene standards are given as their abbreviations in the legend in (A) and are the full-length, PCR-amplified *nifH* genes from *Klebsiella pneumoniae* 342 (Kpn), *Frankia* species CcI3 (Fsp), *Geobacter uraniireducens* RF4 (Gur), *Desulfovibrio vulgaris* Hildenborough (Dvu), and *Mastigocladus laminosus* LB 1931 (Mla).


(A)
Kpn
Fsp
Gur
Dvu
Mla
PolF/PolR
(D)
F2/R6
(B)
nifH1/nifH2
(E)
DVV/IGK3
(C)
Ueda19F/Ueda407R
(F)
DVV/IGK3 I to N
Ct
annealing temperature (°C)

numerically second lowest Ct value for primer sets PolF/PolR and nifH1/nifH2 at the optimal annealing temperature whereas for the primer set Ueda19F/Ueda407R it was numerically highest and for the other primer sets it appeared at an intermediate value.

In our assessment, only the PolF/PolR primer set yielded similar Ct values for each of 5 phylogenetically diverse templates meaning that the amplification efficiencies for each template was similar for this primer set (Figure 5.1). For this primer set the Ct values showed less variation across the range of annealing temperatures tested. To demonstrate, the template *G. uraniireducens* gave a Ct of 10.8 at the annealing optimum of 62 ºC, whereas at 53 ºC the Ct was 15, a difference of 4.2 Ct. However, with the primer set nifH1/nifH2, that same template gave a Ct of 15.7 at the annealing temperature optimum of 60 ºC whereas at 52 ºC the Ct was 22, a difference of 6.3 Ct. In addition, there is a "crossover effect" visible in the Ct vs. annealing temperature curves that is clearest with the primer set nifH1/nifH2, which has several template curves which intersect (Figure 5.1), showing that the relative ratio of the Ct values changes as the annealing temperature changes. This effect is less pronounced for the PolF/PolR primer set (Figure 5.1) with the exception of the *D. vulgaris* template which at lower annealing temperatures crossed over to have a numerically higher Ct value indicating poorer amplification efficiency. For this reason it is important to note that annealing temperature optimization is critical because were we to have carried out the qPCR assay with the PolF/PolR primer set at a lower annealing temperature, like 56 ºC, then *D. vulgaris* would have a Ct of 20, while *G.*

uraniireducens would have a Ct of 13 and the others would have a Ct of 15. At an optimal annealing temperature of 62 ºC, the Ct differences between *D. vulgaris* and other templates are effectively gone.

With the DVV/IGK3 primer set we replaced the primer positions with Inosines (Table 5.2) with N, resulting in templates that are a mixture of oligonucleotides containing all 4 bases at these degenerate positions. Both primers have 5 inosines apiece meaning the modified primers have 10 total positions which contain a mixture of the 4 bases, which effectively decreased the concentration of any one oligonucleotide in the mixture by increasing the total number of oligonucleotides that comprise each primer by 4^5, or 1024 times. This had the effect of numerically increasing the Ct values for the modified primer set and also affecting the position of certain templates relative to others. For example, when the *G. uraniireducens* template was evaluated with the N-primers, the annealing optimum of 56.4 C yielded a Ct of 23.9 whereas with the inosine-primers it yielded a Ct of 19.1 (Figure 5.1F). Also, the *K. pneumoniae* and *D. vulgaris* templates appeared to switch position relative to the other templates (Figure 5.1F). However, replacing Inosines with N did not remedy the template-specific bias effect.

Degenerate primers are often intended to capture the full scope of extant diversity for amplification of a particular gene from an entire microbial community. Based upon our results, and depending upon the composition of the community and the particular standard chosen as a reference, the determined value of the gene copy

number of a community could vary widely. For instance, if a biased primer set like nifH1/nifH2 were employed and *G. uraniireducens* were used as a standard while the microbial community composition was composed primarily of organisms with an amplification efficiency like that of *D. vulgaris*, then the absolute value of the determined copy number would be about 8 copies when there are actually 2.9×10^5 copies present (Figure 5.2). An assay at the optimal annealing temperature of 60 ºC for this primer set would yield a Ct of 16 for the *G. uraniireducens* standard whereas a community consisting of *D. vulgaris* DNA would yield a Ct of 31 for an equivalent copy number. This worst-case scenario of a 15 cycle difference translates to a 2^{15}-fold, or 32,768-fold, underestimate of copy number for the community DNA. In the contrary scenario, use of *D. vulgaris* as the standard for a community consisting of *G. uraniireducens* would give an equivalent overestimate of the actual number of gene copies in the community. Given that microbial community composition can vary dramatically between sites, treatments, and environments, and since most studies quantify gene copy number between samples that would likely have very different community composition, then our results would suggest that there is great potential for mis-estimation within qPCR studies that employ degenerate primers.

application of PolF/PolR to determine *nifH* copy number

Given that the PolF/PolR primer set showed negligible template-specific bias,

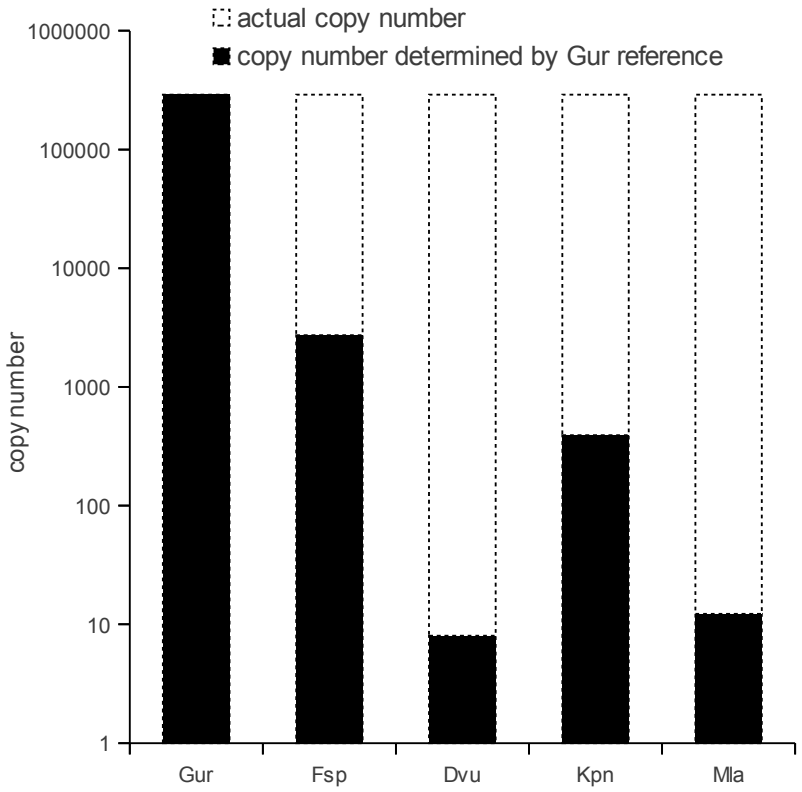
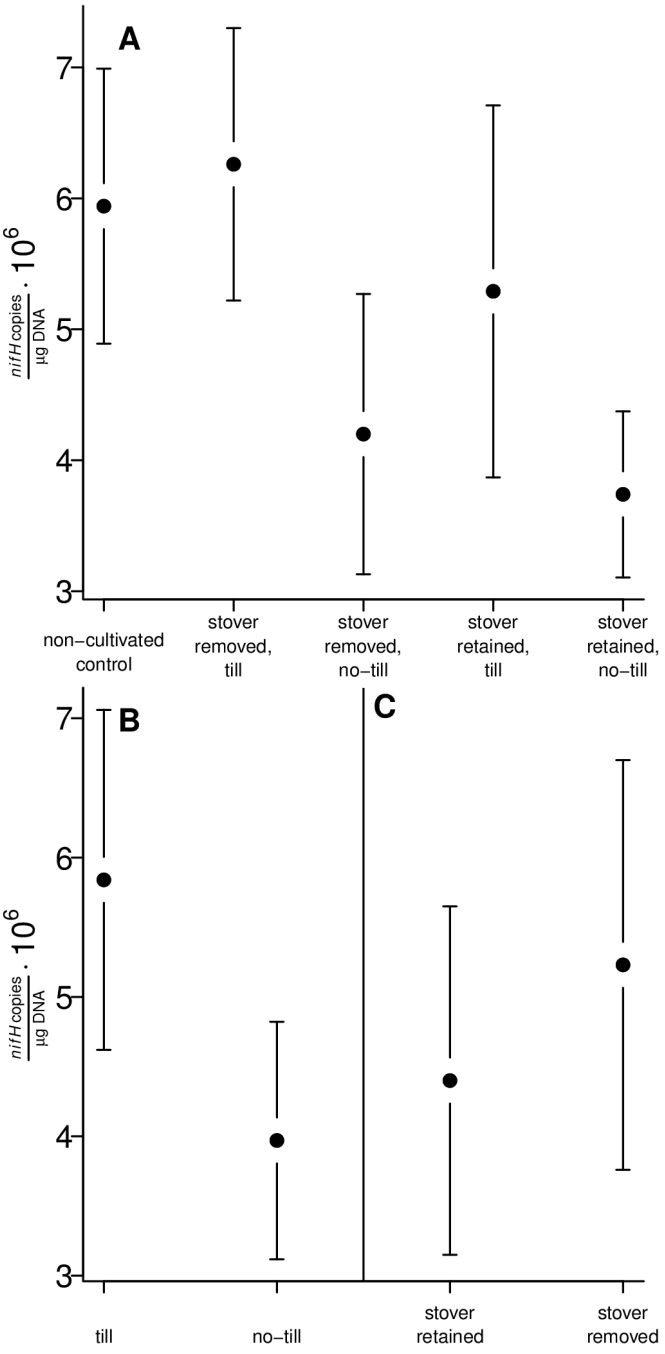


Figure 5.2: Copy number may be dramatically miscalculated when using a biased primer set. Values are taken from the primer set nifH1/nifH2 (Figure 5.1B) and used to calculate effective copy number were the Gur template to be used as a standard. In this case, the 4 other templates, though they are all present at 2.9 x 10^5 copies like Gur,effectively appear to have far fewer copies due to the preferential amplification of Gur which led to the Ct differences seen in Figure 5.1B. In a worst case scenario where the poorly amplified Dvu template constitutes the entirety of a microbial community but Gur is used to make the standard curve, it would appear as if there were only 8 copies present when in fact there were 2.9 x 10^5. The identity of the labels, which represent different *nifH* templates, are given in the legend for Figure 5.1. The identity of the bars is described in the legend at the top of the figure.

we chose to use this primer set to quantify the *nifH* copy number in plots from a long-term agricultural experiment site. The *nifH* gene copy number at the site was 5.43 ± 2.06 x 10^6 (mean ± standard deviation) copies μg^{-1} soil DNA extract with a range from 3.04 x 10^6 to 12.1 x 10^6 copies μg^{-1} soil DNA (Figure 5.3). The highest value was determined to be an outlier by both Grubbs Test (G = 3.0130, p=0.0014) and Dixon Test (Q = 0.6394, p-value = 0.0047). After removing the outlier, which fell into the biomass retained with tillage group, the *nifH* gene copy number at the site was 5.08 ± 1.38 x 10^6 copies μg^{-1} soil extract. With the outlier included, differences between individual treatments were not significant as determined by one way ANOVA (Figure 5.3A). However, with the outlier excluded, there was a significant effect ($F_{3, 11}$=4.65, p = 0.0246; Figure 5.3A). A Dunnett-Tukey-Kramer post hoc test for unbalanced designs due to the excluded outlier showed no significant difference among the 4 treatments. However, imputation followed by Tukey's HSD test showed a significant difference between the biomass removed with tillage and biomass retained without tillage treatments. Significant differences were observed with both the mean substitution (p = 0.0173) approach to imputation and a "hot deck" approach whereby the highest value in that treatment was substituted in place of the outlier (p = 0.0229). Factorial ANOVA revealed that when the outlier was included, the main effect of tillage was significant ($F_{1, 12}$ = 7.24, p = 0.0196; Figure 5.3B) though the effects of biomass retention ($F_{1, 12}$ = 0.018, p = 0.895; Figure 5.3C) and the interaction of biomass retention and tillage ($F_{1, 12}$ = 0.364, p = 0.557) were not significant. When the

Figure 5.3: The *nifH* copy number for the Chazy agricultural site. The dot represents the mean of 4 replicates for each treatment and the error bars are the standard deviation. The *nifH* copy number is shown for the agricultural treatments (A) and the main effects of stover biomass removal vs. retention (B) and till vs. no till (C). An outlier was discarded for the stover retained, till treatment, and thus n = 3 for this treatment whereas for all other treatments n = 4 (A). For the stover retained and till effects n = 7 due to removal of an outlier, but for the other effects n = 8 (B and C).


A
B
C
7
6
5
4
3
nifH copies / µg DNA · 10^6
non−cultivated control
stover removed, till
stover removed, no−till
stover retained, till
stover retained, no−till
till
no−till
stover retained
stover removed

outlier was excluded, the same was true, the main effect of tillage was still significant ($F_{1, 11} = 12.07$, $p = 0.00520$) and the effects of biomass retention ($F_{1,11} = 1.673$, $p=0.222$) and the interaction of biomass retention and tillage ($F_{1,11} = 0.222$, $p = 0.646$) were still not significant. As with the above analyses, $\log_{10}$ transformation of the *nifH* copy number yielded no significant differences between treatments when the outlier was included, but the mean and "hot deck" imputations to replace the outlier showed a significant difference between the biomass removed with tillage and biomass retained without tillage treatments as before.

Discussion

Functional gene qPCR is used widely as a method of measuring the abundance of microbial functional groups in environmental samples. Most universal functional gene primers have significant levels of degeneracy but the effect of primer degeneracy on the validity of qPCR gene copy estimates has been poorly characterized. Our approach to evaluate bias caused by degenerate qPCR primers employed 5 *nifH* gene standards selected to represent the full-range of sequence diversity in *nifH*. We chose PCR-amplified, full-length *nifH* genes as standards rather than genomic DNAs to allow for precise copy number determination in standards, eliminating concerns pertaining to gene dose effects, copy number within the genome, and the impact of non-target sequences in the genome. We evaluated 5 universal *nifH* primer sets which

had different primer binding sites and different degrees of degeneracy over the range of annealing temperatures from 53 to 67 °C. Most primer sets evaluated had substantial amplification bias across the different templates examined (Figure 5.1). The difference, which amounts to greater than 10 cycles between many templates, would result in a greater than 1000-fold difference in the quantity of *nifH* genes determined. In the case of the primer set nifH1/nifH2, the difference between the *nifH* templates from *G. uraniireducens* and from *D. vulgaris* at 60 °C is 15 cycles, equivalent to a 32,768-fold difference (Figure 5.1B). One consequence is that biased primers can generate different gene abundance values for communities that actually have identical gene copy number but vary in community composition.

A more pernicious problem is that the choice of template used to construct a standard curve can cause a systematic and dramatic mis-estimation of gene copy number which can vary from environment to environment as a result of inherent differences in community composition. Such a systematic bias would occur if the template used in the standard curve has a different amplification efficiency from the majority of template found in a given environmental sample. Given the potential for the composition of nitrogen-fixing communities to differ by environment sampled [29, 30] or due to treatment effects [30], the use of an unbiased primer set is critical for obtaining valid results for a quantitative technique.

Our results contrast with earlier results by Zehr and Capone [17] where early amplification competition experiments with *nifH* templates that varied in G-C content

at the primer annealing site did not indicate preferential amplification of templates. Assessment of amplification in this earlier study was based upon interpretation of band intensity, however, and is not relevant to results obtained by qPCR. In addition, Tan et al. [16] used Terminal Restriction Fragment Length Polymorphism (TRFLP) analysis to assess template-specific primer bias using the primer set nifH2/nifH1 with six different templates. The templates were used both individually and in mixture, and the relative template ratios as determined by peak area analysis were maintained. Diallo et al. also used Denaturing Gradient Gel Electrophoresis (DGGE) analysis of the *nifH*, *anfH*, and *vnfH* genes from *Azotobacter vinelandii* and found that the nifH2/nifH1 primer set did not generate bands of different intensities for the 3 gene products [15]. With TRFLP and DGGE, it is the end product of the PCR reaction that is analyzed in contrast to qPCR where product accumulation is observed in real-time. Thus, TRFLP and DGGE would not necessarily reveal differences in amplification efficiency between different templates since these differences may only be visible in the early cycles of PCR and could disappear in later cycles if there is an equilibration in abundance of the different sequence products. Our qPCR analyses indicate substantial bias in amplification efficiency with primers nifH2/nifH1 (Figure 5.1 B). Bias in PCR products would be expected to result from bias in amplification efficiency but analysis of end products is less informative than direct real-time observation for assessing the dynamics of PCR amplification.

The primer sets F2/R6 and nifH2/nifH1 bind within the same template position

as PolF/PolR (Table 5.2), and yet they exhibit considerable primer bias relative to PolF/PolR. F2/R6 contains inosines in place of most degeneracies, whereas nifH2/nifH1 are more degenerate than the other two primer sets. F2/R6 and nifH2/nifH1 are shorter than PolF/PolR and are positioned differently. Since these 3 primer sets are similar in sequence composition but give very different results with regard to bias, this suggests further work to examine the contribution of each degeneracy to the primer bias problem whereby stepwise, single-base modifications are made to a primer and then the primer is examined for a change in performance.

The small difference in mean *nifH* copy number between soils of the agricultural treatments and with the non-cultivated control soil (Figure 5.3) indicate that 30 years of agricultural management had only a modest impact on *nifH* copy number in soils. Within the agricultural site it was clear that tillage caused an increase in *nifH* copy number relative to no-tilled soils, though the effects of biomass retention was not significant. The higher abundance in tilled soils could be due to more favorable conditions for nitrogen fixers under this treatment effect. The tilled soils have overall less organic matter, total C and N, and NO_3 (see summary soil characteristics for the treatments in [32] and are visibly more compacted due to loss of soil aggregate structure to the point where the tilled soils drain poorly. These conditions may have led to more anoxic soil with low N under tillage whereby the ability to fix nitrogen is a selective advantage. Other studies have measured significant differences ($p<0.05$) for a less than 2-fold mean difference in *nifH* copy number

between roots of cucumber under different fertilization levels [33] or for pasture under different stocking rates [33].

Though our work identified the PolF/PolR primer set as having the least template-specific bias, we have not found an unbiased primer with total coverage of all *nifH* genes given that the PolF/PolR primer set has been shown to cover only 25% of *nifH* [2]. However, our work does demonstrate an approach for assessing primer bias and thus represents a way forward. In brief, our approach begins with selection of a set of gene standards which encompass both the full phylogenetic breadth of the gene of study and a fair representation of sequence variation at the site of primer binding. The chosen templates should not have mismatches to the primer sets to be evaluated, and ideally the primers would be minimally degenerate but would still target a majority of the known sequences for the gene, though this is often a trade-off. The full-length gene standards would be amplified by PCR (as opposed to using genomic DNA for which gene copy number concerns arise), and then the amplified product, once purified, would be quantified, then all templates set to an equal copy number. Next, these templates would be evaluated in a standard qPCR reaction against a range of primer sets and over a range of annealing temperatures. Those primers which work well will yield similar Ct values for each template.

Our results suggest a cautious interpretation or reexamination of qPCR studies that employ degenerate primer sets which extends beyond just studies of *nifH*. A solution could be developed by finding a truly universal *nifH* primer set without

amplification bias, by developing several primer sets whose coverage is non-overlapping and altogether encompass the full extent of *nifH* phylogenetic diversity, or by selecting primers from conserved regions of *nifD* or *nifK*. Future work should examine amplification efficiency differences and the potential to account for and correct them, and particular primer or template characteristics which could be responsible for the bias should be identified.

References

1. Moisander PH, Beinart RA, Hewson I, White AE, Johnson KS, Carlson CA et al. (2010) Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. Science 327: 1512-1514.

2. Gaby JC, Buckley DH (2012) A Comprehensive Evaluation of PCR Primers to Amplify the *nifH* Gene of Nitrogenase. PLoS One 7: e42149.

3. Throbäck IN, Enwall K, Jarvis A, Hallin S (2004) Reassessing PCR primers targeting *nirS, nirK* and *nosZ* genes for community surveys of denitrifying bacteria with DGGE. FEMS Microbiol Ecol 49: 401-17.

4. Kanagawa T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). J Biosci Bioeng 96: 317-323.

5. Polz M, Cavanaugh C (1998) Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol 64: 3724-3730.

6. Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K, Nikolausz M (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiol Ecol 60: 341-350.

7. Frank JA, Reich CI, Sharma S, Weisbaum JS, Wilson BA, Olsen GJ (2008) Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. Appl Environ Microbiol 74: 2461-70.

8. Kurata S, Kanagawa T, Magariyama Y, Takatsu K, Yamada K, Yokomaku T et al. (2004) Reevaluation and reduction of a PCR bias caused by reannealing of templates. Appl Environ Microbiol 70: 7545-7549.

9. Bru D, Martin-Laurent F, Philippot L (2008) Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. Appl Environ Microbiol 74: 1660-1663.

10. Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol 62: 625-630.

11. Reysenbach AL, Giver LJ, Wickham GS, Pace NR (1992) differential amplification of ribosomal-RNA genes by polymerase chain-reaction. Appl Environ Microbiol 58: 3417-3418.

12. Ishii K, Fukui M (2001) Optimization of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR. Appl Environ Microbiol 67: 3753-3755.

13. Chandler D, Fredrickson J, Brockman F (1997) Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. Mol Ecol 6: 475-482.

14. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (2005) PCR-induced sequence artifacts and bias: Insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl Environ Microbiol 71: 8966-8969.

15. Demba Diallo M, Reinhold-Hurek B, Hurek T (2008) Evaluation of PCR primers for universal *nifH* gene targeting and for assessment of transcribed *nifH* pools in roots of *Oryza longistaminata* with and without low nitrogen input. FEMS Microbiol Ecol 65: 220-8.

16. Tan Z, Hurek T, Reinhold-Hurek B (2003) Effect of N-fertilization, plant genotype and environmental conditions on *nifH* gene pools in roots of rice. Environ Microbiol 5: 1009-15.

17. Zehr JP, Capone DG (1996) Problems and promises of assaying the genetic potential for nitrogen fixation in the marine environment. Microb Ecol 32: 263-281.

18. Zehr JP, McReynolds LA (1989) Use of degenerate oligonucleotides for amplification of the *nifH* gene from the marine cyanobacterium *Trichodesmium thiebautii*. Appl Environ Microbiol 55: 2522-2526.

19. Mehta MP, Butterfield DA, Baross JA (2003) Phylogenetic diversity of nitrogenase *(nifH)* genes in deep-sea and hydrothermal vent environments of the Juan de Fuca Ridge. Appl Environ Microbiol 69: 960-70.

20. Marusina AI, Boulygina ES, Kuznetsov BB, Tourova TP, Kravchenko IK, Gal'chenko VF (2001) A system of oligonucleotide primers for the amplification of *nifH* genes of different taxonomic groups of prokaryotes. Mikrobiologiya 70: 86-91.

21. Lueders T, Friedrich M (2003) Evaluation of PCR amplification bias by terminal restriction fragment length polymorphism analysis of small-subunit rRNA and mcrA genes by using defined template mixtures of methanogenic pure cultures and soil DNA extracts. Appl Environ Microbiol 69: 320-326.

22. Poly F, Monrozier LJ, Bally R (2001) Improvement in the RFLP procedure for studying the diversity of *nifH* genes in communities of nitrogen fixers in soil. Res Microbiol 152: 95-103.

23. Ueda T, Suga Y, Yahiro N, Matsuguchi T (1995) Remarkable N_2-fixing bacterial diversity detected in rice roots by molecular evolutionary analysis of *nifH* gene sequences. J Bacteriol 177: 1414-7.

24. Ando S, Goto M, Meunchang S, Thongra-ar P, Fujiwara T, Hayashi H et al. (2005) Detection of *nifH* sequences in sugarcane (*Saccharum officinarum* L.) and pineapple (*Ananas comosus* [L.] Merr.). Soil Sci Plant Nutr 51: 303-308.

25. Moebius-Clune BN, van Es HM, Idowu OJ, Schindelbeck RR, Moebius-Clune DJ, Wolfe DW et al. (2008) Long-term effects of harvesting maize stover and tillage on soil quality. Soil Sci Soc Am J 72: 960-969.

26. Wallenstein MD, Vilgalys RJ (2005) Quantitative analyses of nitrogen cycling genes in soils. Pedobiologia 49: 665-672.

27. (2004) Real-time PCR: an essential guide. Horizon Bioscience, Wymondham, Norfolk.

28. Zehr JP, Jenkins BD, Short SM, Steward GF (2003) Nitrogenase gene diversity and microbial community structure: a cross-system comparison. Environ Microbiol 5: 539-54.

29. Gaby JC, Buckley DH (2011) A global census of nitrogenase diversity. Environ Microbiol 13: 1790-9.

30. Culman SW (2008) Soil Microbial Dynamics and Associative Nitrogen Fixation in Kansan Tallgrass Prairies. (Doctoral Dissertation).

31. Hsu S, Buckley DH (2009) Evidence for the functional significance of diazotroph community structure in soil. ISME J 3: 124-36.

32. Juraeva D, George E, Davranov K, Ruppel S (2006) Detection and quantification of the *nifH* gene in shoot and root of cucumber plants. Can J Microbiol 52: 731-9.

33. Wakelin SA, Gregg AL, Simpson RJ, Li GD, Riley IT, McKay AC (2009) Pasture management clearly affects soil microbial community structure and N-cycling bacteria. Pedobiologia 52: 237-251.

CHAPTER 6

AN ECOLOGICAL SURVEY OF NITROGEN-FIXING BACTERIA AND NITROGEN FIXATION IN OLD FIELDS OF NEW YORK STATE

Introduction

While nitrogen is abundant in earth's atmosphere as dinitrogen gas, this chemically stable form of the element is unavailable for uptake by plants which absorb most of their nitrogen as ammonium or nitrate. However, plants do not have the genetic capability of carrying out the enzymatic transformation of dinitrogen gas into ammonia, the process known as Biological Nitrogen Fixation (BNF), and instead must rely upon microorganisms of the domains *Bacteria* and *Archaea* which are the only organisms genetically capable of BNF.

While most inputs from BNF in soils derive from symbiotic associations between *Bacteria* and leguminous plants [1], there exists a vast phylogenetic diversity of nitrogen-fixers that do not form direct symbiotic associations with plants [2]. Much of microbial diversity is represented by organisms that have yet to be grown in culture [3], and similarly many of the surveyed nitrogen-fixers have yet to be isolated in culture and characterized [2, 4]. Cultivation-independent techniques which employ

Polymerase Chain Reaction (PCR) allow us to identify and study the distribution of the many diverse, non-cultivated nitrogen fixers.

Most PCR-based studies of nitrogen fixers make use of the *nifH* gene as a marker. The *nifH* gene is one of three genes that encode the structural components of the nitrogenase enzyme that carries out nitrogen fixation. The *nifH* gene encodes the dinitrogenase reductase which imparts reducing equivalents onto the core of the enzyme complex, thereby driving the reduction of N_2 to NH_3 [5]. In addition to *nifH*, the genes *nifD* and *nifK* encode the core of the enzyme where the actual reduction of nitrogen takes place [5]. The *nifH* gene is the most widely sequenced of the three nitrogenase structural genes due to its higher sequence conservation.

Quantitative, real-time PCR has proven a useful technique for tracking the abundance of non-cultivated organisms in the environment [6]. A number of qPCR studies have examined *nifH* gene abundance to find relationships with chemical, physical, and biological factors. One study reported correlations with electrical conductivity, microbial biomass N, microbial biomass C, total N and total K [7], while another revealed correlation with soil texture showing that clay soils had higher *nifH* abundance than sandy ones and that pH, organic matter, and NH_4+ explained abundance depending on soil texture [8]. In pasture, the amount of N_2 fixed, application of lime, and the stocking rate of livestock affected *nifH* abundance [9]. Cucumber plants showed effects on *nifH* abundance with plant age and nitrogen supply [10], whereas in a sorghum crop, abundance was higher under manure

amendment than under urea and straw, but only at the crop's flowering stage [11]. Plant growth stage was also a significant effect in alfalfa [12]. The amount of nitrogen fertilizer and sorghum cultivar affected *nifH* abundance in the rhizosphere [13], and in a comparison of successional versus agricultural sites organic carbon was a significant effect [14]. Previous studies have shown an absence of correlation of *nifH* gene abundance with land use type [7], phosphorus fertilization in pasture [9], and NPK fertilization and diurnal cycle in rice [15]. However, it is important to note that organisms may possess the genetic capability to carry out nitrogen fixation without ever needing to use it. Thus, factors which increase the absolute abundance of nitrogen-fixing organisms may not actually be selecting for nitrogen fixation itself, but perhaps constrain the abundance of microorganisms in general.

Nitrogen fixation is a biochemical process with certain constraints which may help us make inferences about the ecology of nitrogen-fixing bacteria. The process is energetically demanding [5], and thus nitrogen fixers would expend energy to obtain nitrogen whereas organisms able to scavenge available nitrogen would conserve energy relative to nitrogen fixers and would have a selective advantage. As a consequence of being energetically demanding, nitrogen fixation only occurs under particular environmental conditions. First, nitrogen fixers must be N limited, and nitrogen fixation has been shown to occur in high C:N substrates like dead wood in forests but this required a high moisture content and occurred maximally during summer [48]. Those bacteria carrying out nitrogen fixation require a source of energy

in addition to being N limited, and in the case of soil heterotrophs energy would principally be derived from carbon. Heterotrophic nitrogen fixers in soil need available carbon to drive nitrogen fixation, and the addition of labile carbon as sugars is widely known to stimulate nitrogen fixation [16-18, 19]. Indeed, studies show an increase in nitrogen fixation when organic matter with a high C to N ratio is added to the soil, like the addition of straw has been shown to stimulate nitrogen fixation at field moisture [20, 21], an effect which has also been confirmed in situ [22]. Furthermore, the overlying tree species was shown to control rates of nitrogen fixation in leaf litter and soil in a tropical rain forest, showing that plant species composition may affect nitrogen fixation [49]. Moisture has been shown to impact nitrogen fixation rate [16, 18, 23]. Moisture in environments like soil can control oxygen availability, but more importantly is a control on organismal growth and metabolic activity. Soil pH and the correlated factor Ca were shown to correlate to nitrogen fixation rate, which was also inhibited by low pH but increased up to neutral pH [24]. In a separate study, soil nitrogenase activity was observed to decrease outside the 7-7.5 pH range [25].

Given the lack of knowledge about the ecological constraints on nitrogen fixation [26], our goal with this research was to examine a range of environmental factors which may be expected to influence the rate of nitrogen fixation and the abundance of nitrogen fixers. Our experimental design used 20 sites from two regions of New York State that were selected to be similar in vegetation (old field sites), recent

land use history, and soil texture in order to limit variables and focus more on natural variability in factors like organic matter content, pH, mineral nitrogen content, etc. By looking for regional differences, we could determine if there were measurable effects on nitrogen fixation rate that may have been due to differences in geology, climate, or nitrogen deposition rate. These data point to general factors which may explain the abundance and activity of soil nitrogen fixers in other grassland systems.

Materials and Methods

Site selection and soil sampling

Twenty sites were sampled in October and November, 2008 in two regions of New York State, USA (Table 6.1). The first region is the Fingerlakes region where sites were sampled from Tompkins and Seneca counties, and the second region is the Champlain region where sites were sampled from Clinton and Essex counties. There are 12 sites sampled from the Finger Lakes region and 8 sites sampled from the Champlain region. The 8 Champlain sites were sampled last, on November 22nd and 23rd (Table 6.1). The sites span 390.9 km. All sites were old fields or meadows vegetated mostly by grasses, goldenrod, and aster. The sites were selected to have similar soil textures and little or no slope based upon descriptions in the county soil surveys. The land owner or manager for each site was contacted to determine that no

Table 6.1 Location, soil type, and sampling date for the 20 sampling sites used in this study.

site name[a]	code	latitude[b]	longitude[b]	elevation (m)	soil type[c]	sampling date
Caldwell Field	CW	42.45080	-76.45899	291	Williamson very fine sandy loam	10/20/2008
Monkey Run	MR	42.46969	-76.42992	316	Hudson silty clay loam	10/20/2008
Arnot Forest	AN	42.27620	-76.65824	579	Volusia channery silt loam	10/24/2008
Mount Pleasant	MP	42.46221	-76.37834	504	Erie channery silt loam	10/24/2008
Etna Fringed Gentian Area	EF	42.50034	-76.43357	341	Langford channery silt loam	10/31/2008
Mitchell Street	MT	42.43491	-76.47131	266	Hudson silty clay loam	10/31/2008
North Campus Area	NC	42.46126	-76.47900	279	Unsurveyed Area	11/2/2008
Cascadilla Meadows	CM	42.44396	-76.47129	255	Chenango gravelly loam	11/2/2008
McBride	MB	42.46911	-76.82279	388	Chenango channery silt loam	11/5/2008
Ahouse West	AW	42.59842	-76.8312	327	Conesus gravelly silt loam	11/5/2008
Aurora Musgrave Farm	AM	42.73189	-76.66226	237	Lima silt loam	11/6/2008
field adjacent to Slim Jim Woods	SJ	42.45166	-76.46300	260	Chenango gravelly loam	11/6/2008
Blanchard	BC	44.32398	-73.44601	70	Pootatuck fine sandy loam	11/22/2008
Ivy	VY	44.30292	-73.46604	163	Georgia loam	11/22/2008
Uihlein Farm	LN	44.23960	-73.98908	611	Monadnock fine sandy loam	11/22/2008
chimney at Uihlein Sugarbush	CN	44.26234	-73.97358	554	Monadnock fine sandy loam	11/22/2008
Willsboro Cabin	WC	44.38098	-73.37504	46	Bombay gravelly loam	11/23/2008
Willsboro Bio	WB	44.38477	-73.38782	59	Bombay gravelly loam	11/23/2008
Rover Farm	RV	44.89690	-73.42465	33	Fluvaquents-Udifluvents complex	11/23/2008
Chazy	CZ	44.88467	-73.47442	57	Roundabout silt loam	11/23/2008

[a] Although some site names say "farm" or "forest" all sites are located within old fields.
[b] Latitude and longitude are in decimal degrees.
[c] Soil types are from the designations in the USDA-NRCS soil taxonomy.

cultivation or grazing had occurred within the last 5 years. The sites, which occur where forest is the climax community, were maintained as grassland mainly though mowing. Four aggregate soil samples were collected from each site for testing. At each site, 4 parallel transects were demarcated, each 20 meters long and 10 meters apart. Ten soil cores 2.5 cm in diameter and approximately 5 cm in depth were taken along the full length of each transect, and the cores from each transect were sieved with a 4 mm sieve to remove roots and large rocks and to homogenize. A small portion of soil

was flash frozen from the homogenized soils within an hour of sampling with liquid nitrogen and maintained frozen on dry ice until freezer storage in the laboratory. The remaining soil was saved for performing analyses described below. Thus, the 4 replicate transects taken at the 20 sites yields 80 samples for which data were determined.

Nitrogen fixation rate assays

Assays of potential nitrogen fixation rate were conducted within 2 days of sampling. For the assays, 5 g of field moist soil was placed into 18 x 150 mm Balch tubes (Bellco Glass, Vineland, NJ, USA) and incubated in parallel with either unlabeled or ^{15}N-labeled N_2. The tubes were stoppered, sparged 3 times with He, and then the gas inside the tubes was replaced with 80% N_2 (either $^{14}N_2$ or $^{15}N_2$ [>98 atom % ^{15}N, Isotec, Miamisburg, OH, USA]) and 20% O_2. Tubes were incubated for 9 days at room temperature in the dark. Following incubation the soil was homogenized by cryogrinding using a Spex Certiprep 6750 Freezer/Mill. The ^{15}N enrichment of soil samples was determined at the Cornell University Stable Isotope Laboratory by performing Isotope Ratio Mass Spectrometry using a Finnigan MAT Delta Plus mass spectrometer (Thermo Electron Corporation, Waltham, MA, USA) connected to a Carlo Erba NC2500 elemental analyzer (CE Instruments, Wigan, UK) through a Conflo II open split interface (Thermo Electron Corporation). The nitrogen fixation

rate was calculated by determining the amount of N fixed from the ^{15}N enrichment of samples incubated with $^{15}N_2$ relative to unlabeled reference samples incubated in parallel, then dividing this amount by the number of days the soil was incubated and the dry weight of the soil.

Soil variables

A range of soil characteristics were measured for all field samples. Soil moisture was measured gravimetrically by weighing approximately 10 g of field moist soil, then drying the soil in an oven. P_w was then determined as the weight of water contained in the soil divided by the dry weight of the soil. The Bulk density was measured by determining the mass of a fixed volume of dry soil, and then calculated as the weight of dry soil divided by the volume occupied by the soil in the original core. Bulk density values are the average of 3 cores taken for each transect. Percent water-filled pore space (%WFP) soil was determined as $\%WFP = [P_w \times (D_b/S_t)] \times 100$, with total porosity defined as $S_t = 1 –$ (bulk density / particle density) x 100, and particle density estimated as 2.65 g cm^{-3} [27].

Determination of Morgan extractable ions P, K, Mg, Ca, Fe, Al, Mn, Zn, Cu, and NO_3 as well as pH, exchange acidity, and organic matter content via loss on ignition were determined by the Cornell Nutrient Analysis Laboratory using standard methods as defined by the National Soil Survey Center [28] and the Recommended

Soil Testing Procedures for the Northeastern United States [29].

The NO_3 and NH_4 was extracted in 2 M KCl within 2 days of sampling. Briefly, 10 g of field moist soil was extracted with 100 ml 2 M KCl for 1 hour on a rotary shaker. Samples were stored for up to several weeks at 4 °C until processing. Quantification of NO_3 was done by the Greiss-Illosvay technique and NH_4 by the Berthelot or Indophenol Blue method [30]. NO_3 was first reduced to NO_2 using a copperized cadmium column prior to the assay [30]

Potential nitrogen mineralization rates were determined following the Long-Term Ecological Research site protocol [31]. Briefly, soils were adjusted to 60% water-filled pore space and incubated for 28 days at 25 ºC. NO_3 and NH_4 was then extracted with 2 M KCl and the concentration of NO_3 and NH_4 determined as described above. Nitrogen mineralization was determined by subtracting the starting amount of NO_3 and NH_4 in the soil from the final amount of NO_3 and NH_4 at the end of the incubation, then dividing by the 28 days of incubation and the soil dry weight.

qPCR assays

DNA was extracted from 0.25 g of soil using a Powersoil DNA Isolation Kit according to the manufacturer's protocol (MoBio, Carlsbad, CA). Bead-beating was performed using a Mini BeadBeater-8 (Biospec Products, Bartlesville, OK) for two minutes on the homogenize setting. DNA samples were diluted 1:100 for the qPCR

assays which we found to be sufficient to eliminate inhibition effects commonly noted in soil DNA extracts.

For the universal *nifH* qPCR assays, we used the HPLC-purified primers PolF with sequence 5'- TGC GAY CCS AAR GCB GAC TC -3' and PolR with sequence 5'- ATS GCC ATC ATY TCR CCG GA -3' [32]. The qPCR reaction components used for 25 µL reactions were Qiagen Quantifast SYBR Green mastermix, 300 nM each primer, and 5 µl of the DNA template. A 2-step cycling protocol was used on a BioRad Cfx96 qPCR machine with C1000 thermocycler, and cycling parameters consisted of an initial hot-start at 95 °C for 10 minutes, followed by 40 cycles of 95 °C for 15 s and the 62 °C annealing temperature for 60 s. On each plate, 3 qPCR reactions were run for each transect DNA sample (triplicate technical replicates), and the mean of the 3 was then used for calculation of copy number. No-Template Controls (NTC) and at least three dilutions of a *nifH* gene standard of known copy number (as determined by a picogreen DNA quantification assay) were included on each plate. The *nifH* gene standard used was the PCR-amplified, full-length *nifH* gene from *Klebsiella pneumoniae* strain 342. Melt curves showed the samples to have similar profiles to that of the standard included on the plates.

The *nifH* copy number was normalized both with respect to dry weight of soil and then with respect to mass of extracted DNA. The former value represents the total *nifH* gene copy number in soil while the latter represents *nifH* relative abundance in the microbial community.

Statistical analyses

Statistical analyses were performed with the R statistical software (version 2.13.1, The R Foundation for Statistical Computing, http://www.R-project.org). The Welch two sample t-test was used to evaluate the significance of pairwise comparisons. Statistical outliers were identified using Grubbs' test [33]. Where outliers were detected analyses were performed and summary statistics presented both with and without the outliers removed. For the analyses involving testing or comparison of groups defined according to their C and N isotopic enrichment, outliers in the $\delta\ ^{13}C$ profile that did not fall into the one of two distinct modes were excluded along with all 4 samples from the RV site, whose $\delta\ ^{15}N$ signature was enriched due to manure storage nearby, and all 4 samples from the CW site, which had been an agricultural field and whose depleted $\delta\ ^{15}N$ values indicated the influence of fertilizer nitrogen. To eliminate strong positive skew, nitrogen fixation and nitrogen mineralization rates were $\log_{10}$ transformed. To eliminate negative values prior to transformation, 1.0899 was added to the values for nitrogen fixation rate, and 1.07 to those for nitrogen mineralization. Bivariate correlations are reported as the R^2 and p-value for the linear regressions.

Results

Table 6.2 Soil variables measured at each site.

site[a]	*nifH* (10^7 copies g^{-1} dry soil)	*nifH* (10^6 copies μg^{-1} DNA)	soil DNA ($\mu g\ g^{-1}$ dry soil)	N fix. rate ($\mu g\ kg^{-1}\ day^{-1}$)	N min. ($\mu g\ g^{-1}$ dry soil day^{-1})	$\delta^{13}C$ ‰	$\delta^{15}N$ ‰	pC (%)	pN (%)
AM	2.7±1.3	2.4±1.1	11.0±1.7	0.35±0.06	0.49±0.12	-25.0±1.1	5.7±0.6	5.34±0.44	0.49±0.03
AN	4.2±1.8	5.4±1.5	7.4±1.5	0.85±0.13	-0.04±0.05	-27.3±0.1	2.7±0.3	5.57±0.35	0.45±0.03
AW	NA	NA	NA	0.78±0.56	0.72±0.17	-27.3±0.1	3.1±0.3	3.19±0.24	0.28±0.02
BC	0.91±0.1	2.1±0.3	4.3±1.0	0.75±0.96	2.14±0.23	-26.1±0.2	4.6±0.3	1.58±0.25	0.14±0.02
CM	2.9±0.4	3.3±0.5	9.0±1.3	0.39±0.40	1.74±0.44	-24.7±2.7	2.5±0.4	2.97±0.50	0.23±0.03
CN	3.2±0.5	4.3±0.6	7.6±1.6	0.42±0.15	3.06±1.51	-26.4±0.1	4.5±0.6	7.75±1.90	0.59±0.15
CW	2.4±0.3	3.6±0.7	7.0±2.3	0.11±0.03	4.04±0.74	-28.0±0.1	0.7±0.2	3.21±0.16	0.30±0.01
CZ	6.5±1.4	7.5±1.6	8.7±1.3	1.76±1.44	5.38±0.59	-27.6±0.2	4.3±0.2	3.88±0.74	0.36±0.07
EF	3.6±0.5	6.9±0.4	5.2±0.8	0.82±0.45	0.00±0.09	-27.5±0.1	4.0±0.1	3.46±0.14	0.30±0.01
LN	3.5±0.6	3.7±0.4	9.8±2.2	0.51±0.36	8.44±0.64	-26.2±0.1	4.8±0.4	9.21±0.78	0.71±0.05
MB	2.0±0.4	3.7±0.4	5.6±1.4	0.35±0.08	2.28±0.59	-26.1±0.1	3.7±0.1	4.21±0.21	0.38±0.01
MP	7.7±1.8	8.3±1.2	9.7±3.2	0.40±0.06	3.18±0.26	-27.6±0.1	4.6±0.1	5.07±0.30	0.50±0.02
MR	5.9±0.8	7.6±0.6	7.8±1.3	0.27±0.04	3.22±0.64	-27.4±0.2	3.8±0.3	4.99±0.14	0.45±0.01
MT	5.2±1.1	6.2±0.6	8.4±2.7	0.38±0.10	0.87±0.46	-27.5±0.0	2.5±0.3	3.46±0.28	0.30±0.02
NC	1.4±0.3	2.4±0.3	5.7±1.0	0.31±0.08	1.47±0.05	-27.1±0.1	2.3±0.3	2.00±0.15	0.20±0.02
RV	7.6±2.4	6.1±2.5	12.8±2.5	0.27±0.04	6.27±0.75	-27.3±0.1	7.1±0.1	4.85±0.48	0.45±0.02
SJ	1.3±0.2	2.4±0.4	5.3±0.9	0.17±0.11	1.76±0.10	-26.7±0.1	4.9±0.1	3.47±0.14	0.34±0.01
VY	5.7±2.6	8.6±2.8	6.8±3.1	0.45±0.16	2.04±0.60	-27.6±0.1	3.3±0.4	3.71±0.25	0.32±0.02
WB	4.9±2.1	9.2±1.6	5.1±1.6	1.36±1.44	1.21±0.67	-27.6±0.2	2.6±0.1	3.79±0.21	0.31±0.02
WC	3.6±1.1	9.1±1.7	4.0±1.2	0.31±0.06	0.98±0.34	-27.4±0.2	3.0±0.2	3.72±0.29	0.30±0.02

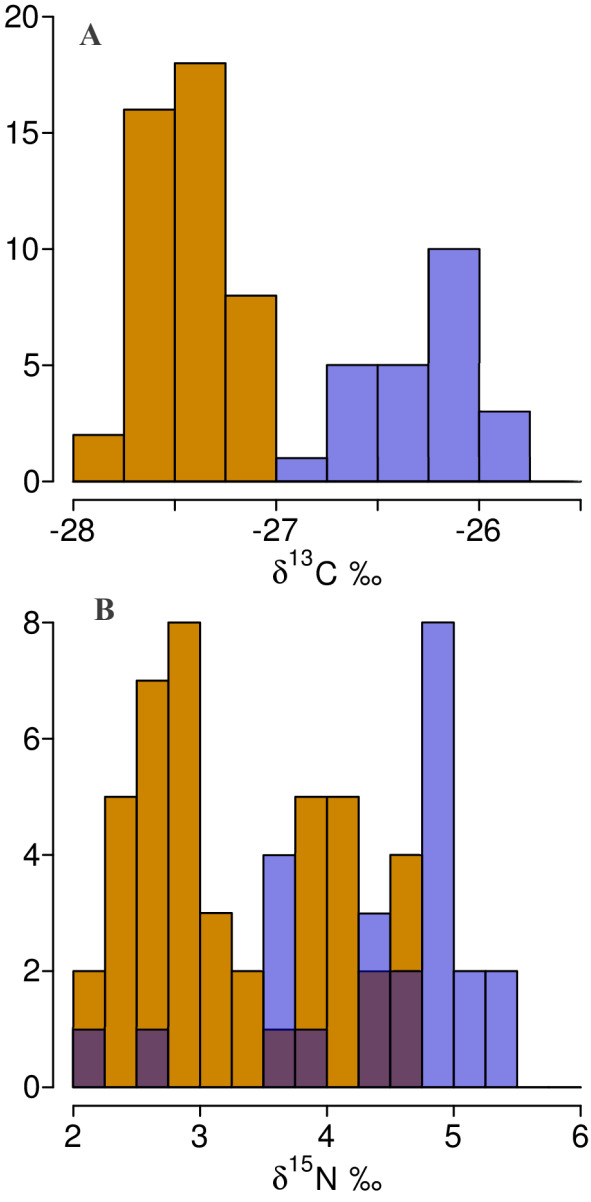
[a] sites are described in Table 6.1. Values are mean ± standard deviation.

Table 6.2 continued

site[a]	C:N	P (kg ha^{-1})	K (kg ha^{-1})	Ca (kg ha^{-1})	Mg (kg ha^{-1})	pH	Pw (%)	LOI (%)	NH_4 (µg g^{-1} dry soil)	NO_3 (µg g^{-1} dry soil)
AM	10.8±0.3	5±1	301±111	7700±791	1257±48	7.2±0.2	42.2±1.8	10.1±0.5	12.3±0.7	5.4±0.5
AN	12.4±0.4	3±1	284±60	2480±274	188±29	5.6±0.2	52.9±5.4	12.0±0.7	9.9±3.3	-0.1±0.1
AW	11.2±0.1	3±1	213±19	5130±319	545±26	6.2±0.2	32.6±1.8	7.1±0.6	9.1±0.7	5.1±2.7
BC	11.6±0.1	18±4	202±15	1140±133	150±7	5.7±0.2	13.7±2.0	2.8±0.6	1.61±0.2	4.8±0.6
CM	13.1±1.5	42±25	325±142	10090±4990	450±65	7.1±0.3	25.1±2.9	5.8±1.1	9.1±1.5	5.9±1.7
CN	13.2±0.3	5±1	90±14	2260±332	154±24	5.5±0.1	44.0±1.5	12.9±0.5	2.3±0.4	5.2±4.7
CW	10.8±0.2	5±1	392±43	2710±99	289±26	5.7±0.1	26.5±1.5	7.4±0.4	10.3±5.4	1.0±0.3
CZ	10.9±0.2	5±3	156±28	5400±852	594±161	6.5±0.4	36.1±3.1	7.6±1.2	3.9±0.4	24.2±10.2
EF	11.7±0.2	1±2	150±22	2610±976	324±73	5.6±0.4	42.7±1.2	7.6±0.2	7.4±3.0	0.5±1.2
LN	12.9±0.3	9±6	272±67	3440±365	440±157	5.4±0.1	55.6±4.6	16.8±1.3	10.4±4.7	17.2±2.8
MB	11.0±0.2	20±2	672±72	1840±222	335±25	5.1±0.1	37.4±2.1	8.7±0.3	8.0±0.8	8.6±3.0
MP	10.2±0.1	6±2	272±36	5180±531	493±48	5.7±0.1	69.8±11.5	11.5±0.8	7.9±3.0	15.1±3.1
MR	11.0±0.1	4±1	256±24	4960±739	574±58	5.9±0.3	40.7±1.6	11.4±0.4	4.5±0.9	0.6±0.3
MT	11.6±0.2	2±0	214±28	3170±367	548±154	5.6±0.1	37.8±1.4	7.4±0.6	13.7±3.9	2.9±2.8
NC	10.2±0.2	5±1	163±8	925±163	115±10	4.9±0.1	19.4±0.6	5.3±1.7	5.0±0.4	6.2±2.6
RV	10.7±0.5	31±4	69±8	6790±429	976±113	6.4±0.4	44.3±6.3	9.0±0.3	2.9±0.3	37.4±9.8
SJ	10.1±0.3	17±1	684±42	1940±107	214±18	5.5±0.1	31.5±0.2	7.3±0.3	6.1±0.5	7.2±0.9
VY	11.8±0.1	2±0	62±8	3270±454	485±96	5.9±0.4	45.2±4.2	7.4±0.5	2.3±0.4	8.3±4.7
WB	12.2±0.4	7±2	126±11	3810±563	482±133	6.4±0.5	36.8±7.8	6.4±0.3	3.0±0.5	11.0±8.2
WC	12.2±0.3	3±0	115±14	3200±503	360±49	5.5±0.2	37.1±4.7	7.2±0.8	3.5±0.7	1.9±2.8

[a] sites are described in Table 6.1. Values are mean ± standard deviation.

Figure 6.1 Histograms of the $\delta^{13}C$ (A) and $\delta^{15}N$ (B) values from the old field sites. The $\delta^{13}C$ values are clearly bimodal. Brown represents the $\delta^{13}C$ depleted group, blue the $\delta^{13}C$ enriched group, and purple the overlap of the two groups in both (A) and (B) showing that the $\delta^{13}C$ depleted group corresponds to the depleted group of the $\delta^{15}N$ values, whose values closer to 0 suggest a greater contribution from nitrogen fixation. Outliers were removed as described in the materials and methods (n=68).


A
20
15
10
5
0
-28
-27
-26
δ13C ‰
B
8
6
4
2
0
2
3
4
5
6
δ15N ‰

Soils from sites in the Finger Lakes and Champlain regions (Table 6.2) differed in NH_4 ($t = -6.6193$, $df = 72.042$, p-value = 5.511×10^{-9}), soil NO_3 ($t = 3.786$, $df = 36.301$, p-value = 0.0005554), nitrogen mineralization rate ($t = 4.0396$, $df = 41.469$, p-value = 0.0002262), $\delta\ ^{15}N$ ($t = 2.8745$, $df = 66.49$, p-value = 0.005429), C to N ratio ($t = 3.6657$, $df = 71.496$, p-value = 0.0004706), and *nifH* relative copy number ($t = 2.5269$, $df = 55.358$, p-value = 0.01439) and these results were significant. In contrast, *nifH* total copy number ($t = 1.6976$, $df = 61.058$, p-value = 0.09468), nitrogen fixation rate ($t = 1.8202$, $df = 36.629$, p-value = 0.0769), $\delta\ ^{13}C$ ($t = -0.8641$, $df = 74.284$, p-value = 0.3903), and pH ($t = 0.37$, $df = 77.598$, p-value = 0.7124) did not differ with respect to region.

The $\delta\ ^{13}C$ values of soil samples revealed a bimodal distribution (Figure 6.1A). We define two groups based upon the $\delta\ ^{13}C$ values, a depleted group with values less than -27 ‰ and an enriched group with values greater than -27 ‰. When we demarcated the $\delta\ ^{13}C$ depleted and enriched groups on the distribution of $\delta\ ^{15}N$ values, we found that the depleted $\delta\ ^{13}C$ values corresponded to depleted $\delta\ ^{15}N$ (Figure 6.1B). Taken together, the values for isotopic enrichment show distinct separation among the two groups (Figure 6.2). We subsequently investigated whether nitrogen fixation and the abundance of nitrogen fixers varied between these two groups of sites, as defined by their isotopic signature.

Nitrogen fixation appeared to predominate in the $\delta\ ^{13}C$ depleted group, whereas nitrogen mineralization appeared to dominate in the enriched $\delta\ ^{13}C$ group. Nitrogen

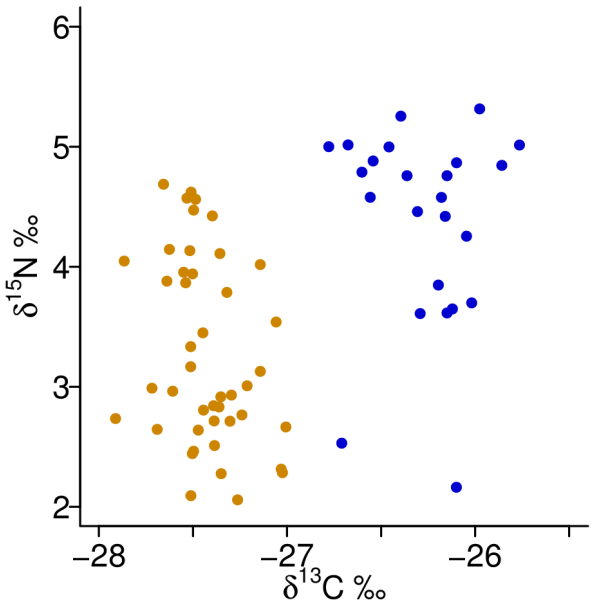
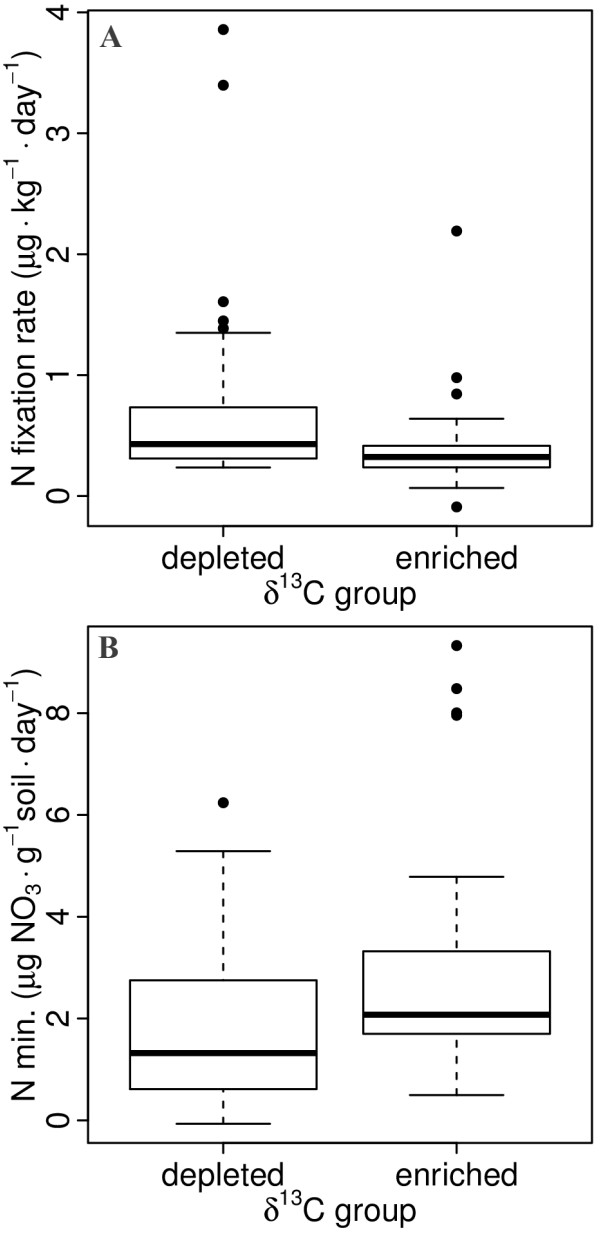


Figure 6.2 δ13C vs. δ 15N values reveal two distinct groups. One group shows enrichment for both isotopes whereas the other shows depletion. The R^2 for the association is 0.1907 with p = 0.0001167. Outliers were removed as described in the materials and methods (n=68).

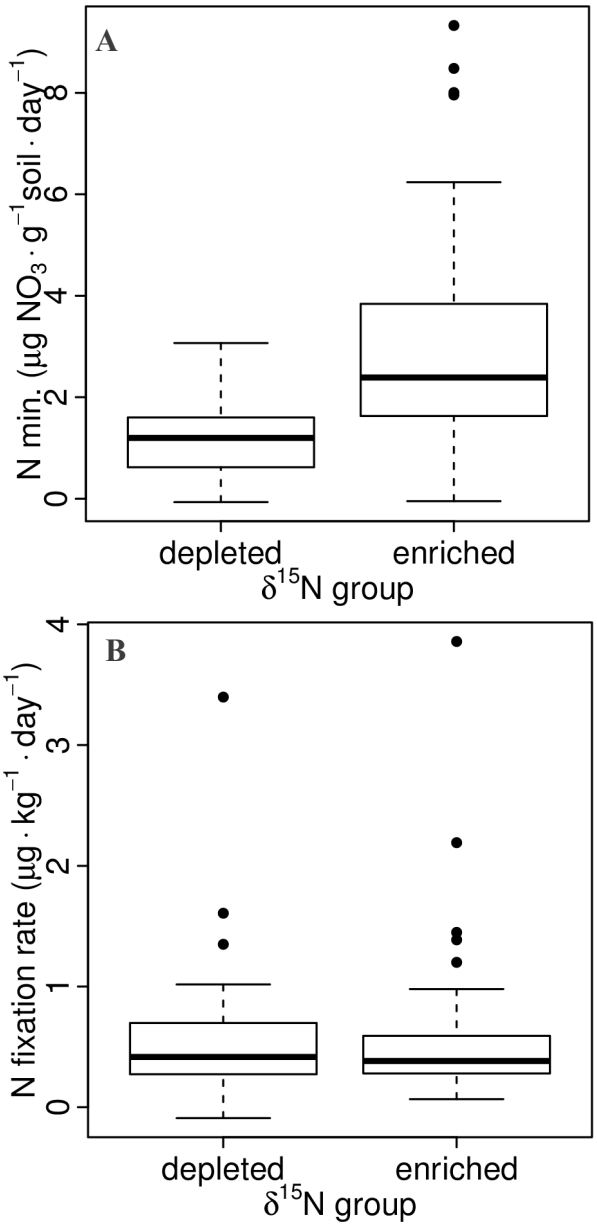
Figure 6.3 Nitrogen fixation rates from the depleted and enriched δ13C groups (Figure 6.1). Nitrogen fixation rates (A) differ significantly but nitrogen mineralization rates do not as detailed in the results. Outliers were removed as described in the materials and methods.


A
N fixation rate (µg · kg^{-1} · day^{-1})
0
1
2
3
4
depleted
enriched
δ^{13}C group
B
N min. (µg NO$_3$ · g^{-1}soil · day^{-1})
0
2
4
6
8
depleted
enriched
δ^{13}C group

fixation rates differed for the two $\delta^{13}C$ groups ($t = 2.184$, $df = 58.215$, p-value = 0.03301; Figure 6.3A). The depleted $\delta^{13}C$ values had the highest rates of potential nitrogen fixation at 0.6994 ± 0.7399 (mean ± s.d.) $\mu g\ kg^{-1}\ day^{-1}$ whereas the enriched ones had a mean of $0.4269 \pm 0.4404\ \mu g\ kg^{-1}\ day^{-1}$. Nitrogen mineralization rates differed between the two isotopically distinct soils with the depleted group having a mean of $1.729 \pm 1.614\ \mu g\ NO_3\ g^{-1}$ soil day^{-1} and the enriched group having a mean of $3.143 \pm 2.582\ \mu g\ NO_3\ g^{-1}$ soil day^{-1} and this result was significant ($t = 3.1282$, $df = 51.61$, p-value = 0.00289; Figure 6.3B). Furthermore, when $\delta^{15}N$ enriched and depleted groups are considered (separated at 3.5 ‰), nitrogen mineralization rates differed significantly ($t = -4.0942$, $df = 64.732$, p-value = 0.0001199; Figure 6.4A) while nitrogen fixation rates did not differ ($t = 0.2182$, $df = 60.713$, p-value = 0.828; Figure 6.4B). Nitrogen mineralization and nitrogen fixation are not positively correlated (Figure 6.5). The nitrogen mineralization rate corresponds weakly to the presence of organic matter in the soil ($p=1.715 \times 10^{-6}$, $R^2= 0.2842$). However, soil organic matter and soil moisture were correlated with a Pearson correlation coefficient of 0.76. Soil 2M KCl extractable NH_4 did not differ between $\delta^{13}C$ groups ($t = -0.7227$, $df = 54.366$, p-value = 0.473), nor did NO_3 ($t = 0.5465$, $df = 73.592$, p-value = 0.5864), nor did NO_3 and NH_4 combined ($t = 0.1522$, $df = 74.717$, p-value = 0.8794).

Assessment of *nifH* relative copy number provides additional evidence that the $\delta^{13}C$ depleted group is undergoing nitrogen fixation while the enriched group group is undergoing nitrogen mineralization. The *nifH* abundance is higher in the depleted

Figure 6.4 The isotopically distinct δ ^{15}N groups have significantly higher rates of nitrogen mineralization (A) but nitrogen fixation rates do not differ (B), as described in the results. Outliers were removed as described in the materials and methods.


A
N min. (µg $NO_3 \cdot g^{-1}$soil $\cdot$ day^{-1})
0
2
4
6
8
depleted
enriched
δ^{15}N group
B
N fixation rate (µg $\cdot$ kg^{-1} $\cdot$ day^{-1})
0
1
2
3
4
depleted
enriched
δ^{15}N group

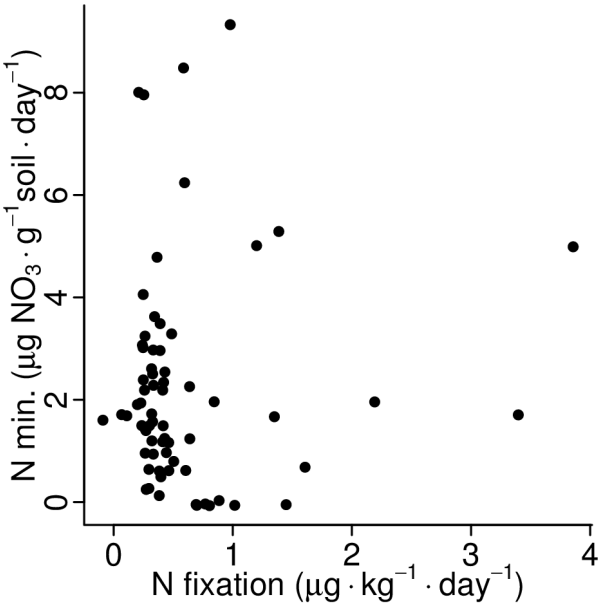


Figure 6.5 Nitrogen fixation and nitrogen mineralization do not correlate. The lack of co-occurrence of high values for both processes in the same site suggests the two processes are not positively correlated. Outliers were removed as described in the materials and methods.

group, and the difference is significant (t = 9.5295, df = 56.843, p-value = $2.193e^{-13}$; Figure 6.6), though this does not hold for the δ ^{15}N groups (t = 1.9219, df = 52.914, p-value = 0.06002). In contrast, high rates of nitrogen mineralization correspond to constrained *nifH* abundance (Figure 6.7). For the biomass-related variables P and K, high values of either correspond to low *nifH* relative abundance, and the high values correspond distinctly to the group with enriched isotope values (Figure 6.8A and B).

Additional factors were observed to correlate with *nifH* abundance. Moisture showed significant correlation to the total and relative abundance of *nifH* ($p = 2.367 \times 10^{-5}$, $R^2 = 0.2077$; Figure 6.9). Ca++ and Mg++ correlated strongly to total *nifH* (Figure 6.10). These variables are themselves highly correlated to pH (Pearson

correlation coefficient = 0.82 for Ca++ and 0.72 for Mg++, n=80; Figure 6.11). Values for pH ranged from 4.9 at the NC site to 7.5 at the AM site. The amount of soil DNA extracted correlated to *nifH* abundance ($R^2 = 0.3039$, $p = 1.481x10^{-7}$; Figure 6.12). The amount of DNA extracted from the soil also correlated with pH, Ca++ ($R^2= 0.3014$, $p=1.696x10^{-7}$; Figure 6.13), as it did with Mg++. Pearson correlation coefficient values for recovered DNA were 0.43 with pH, 0.54 with Ca, and 0.56 with Mg, and with additional variables was 0.41 with Pw, 0.54 with NO_3, and 0.41 with organic matter by LOI.

Discussion

We conducted an ecological survey of soils in order to better understand the factors that influence nitrogen fixation and the abundance of nitrogen fixers. We selected 20 old field and meadow sites in two regions of New York State which were similar in vegetation cover, soil texture, and recent management history in order to reduce variation due to edaphic factors. We measured soil variables (Table 6.2) expected to impact nitrogen fixation based upon ecological or biochemical knowledge [26, 34, 35]. The factors included the availability of nitrogen which was measured through extractable inorganic N assays and N-mineralization determinations, carbon to nitrogen ratio and isotopic signatures, and soil organic matter content. We also

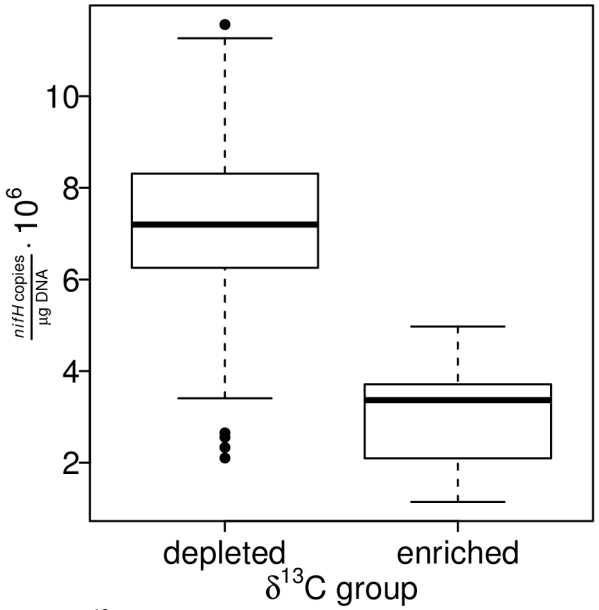


Figure 6.6 The depleted $\delta^{13}C$ group has a higher *nifH* abundance. The outliers in the $\delta^{13}C$ and $\delta^{15}N$ values were excluded as described in the methods.

obtained data for variables which may serve as general indicators of soil fertility like P and K as well as for the general biological controls pH and soil moisture content. An additional consideration we made was the possibility of a regional effect which could have been attributable to geological, climate, or nitrogen deposition differences, but the data did not support regional differences in nitrogen fixation and the significant difference in *nifH* abundance depended on the applied normalization.

The majority of grassland nitrogen fixation is estimated to derive from non-symbiotic fixation given that legumes constitute only a fraction of the primary productivity in grasslands [1]. There is an estimated 2.2 kg N ha^{-1} yr^{-1} non-symbiotically fixed in grasslands though estimates can range as high as about 10 times

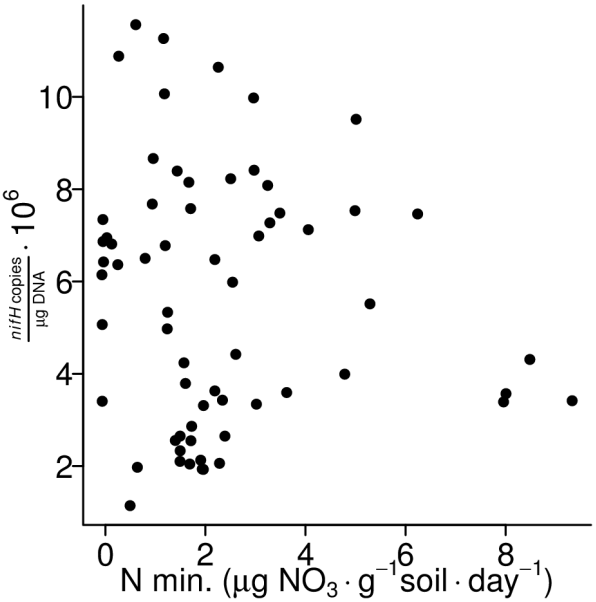
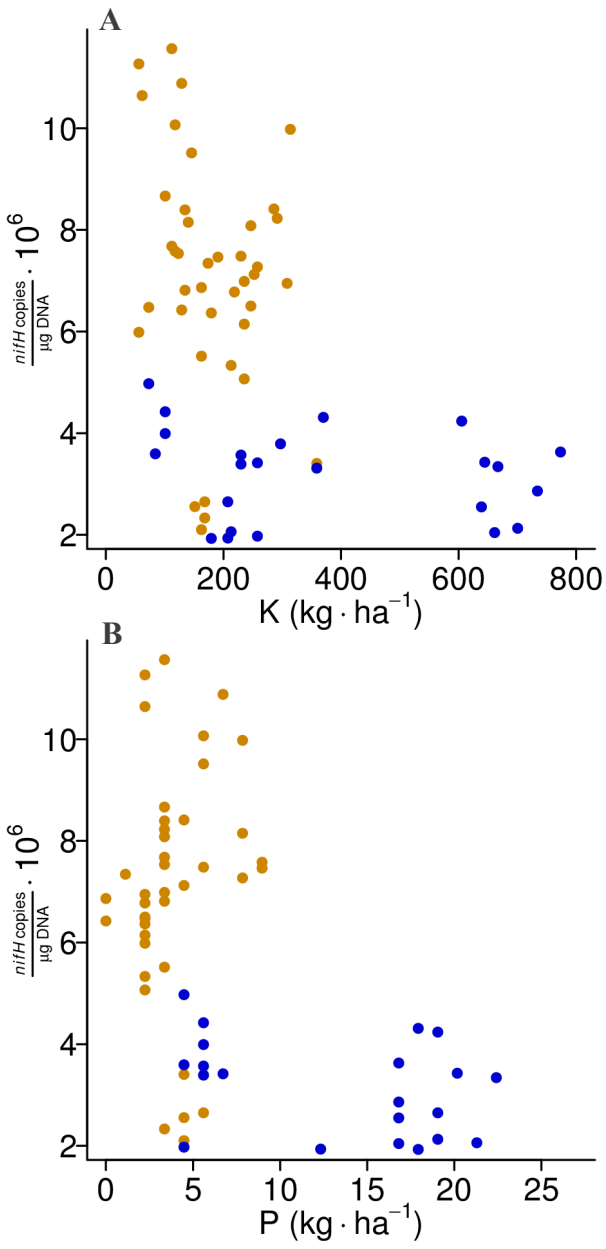


Figure 6.7 Nitrogen mineralization versus with relative *nifH* abundance. High rates of nitrogen mineralization do not co-occur with high *nifH* abundance. Outliers were removed as described in the materials and methods.

that amount [1]. When we scaled up our 0.55 µg kg^{-1} day^{-1} average value for fixation across all sites by assuming a bulk density of 1.65 g cm^{-3} and by assuming the rate measured in the top 5 cm to be constant to 30 cm depth, we determined the nitrogen fixation rate to be 0.5 kg N ha^{-1} 180 $days^{-1}$. However, organic matter content of the soil and other factors which may drive nitrogen fixation likely diminish with increasing depth in the soil profile. Our results are not that different from the 2.2 kg N ha^{-1} yr^{-1} reported elsewhere as an average nitrogen fixation in temperate grasslands [1]. Old field sites in Oklahoma were observed to have rates of nitrogen fixation reduced to one third that of prairie, an occurrence which the authors attributed to the presence of plant species with allelopathic affects [46]. The rates of temperate grassland nitrogen

Figure 6.8 The relationship of K and P to relative *nifH* abundance. Low values of K and P have higher abundance of *nifH*, though constrained samples occur. The outliers in the δ ^{13}C and δ ^{15}N values were excluded as described in the methods. Colors correspond to the δ ^{13}C depleted and enriched groups that are identically colored in Figure 6.1.


A
10
8
6
4
2
0
200
400
600
800
K (kg · ha^{-1})
nifH copies / µg DNA · 10^6
B
10
8
6
4
2
0
5
10
15
20
25
P (kg · ha^{-1})
nifH copies / µg DNA · 10^6

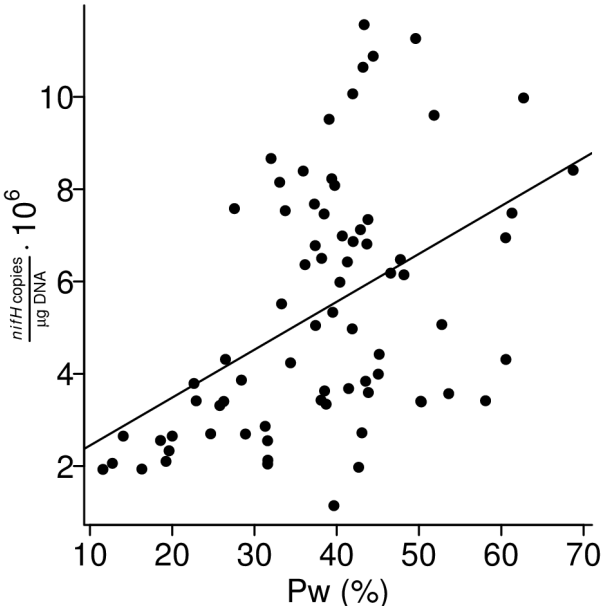
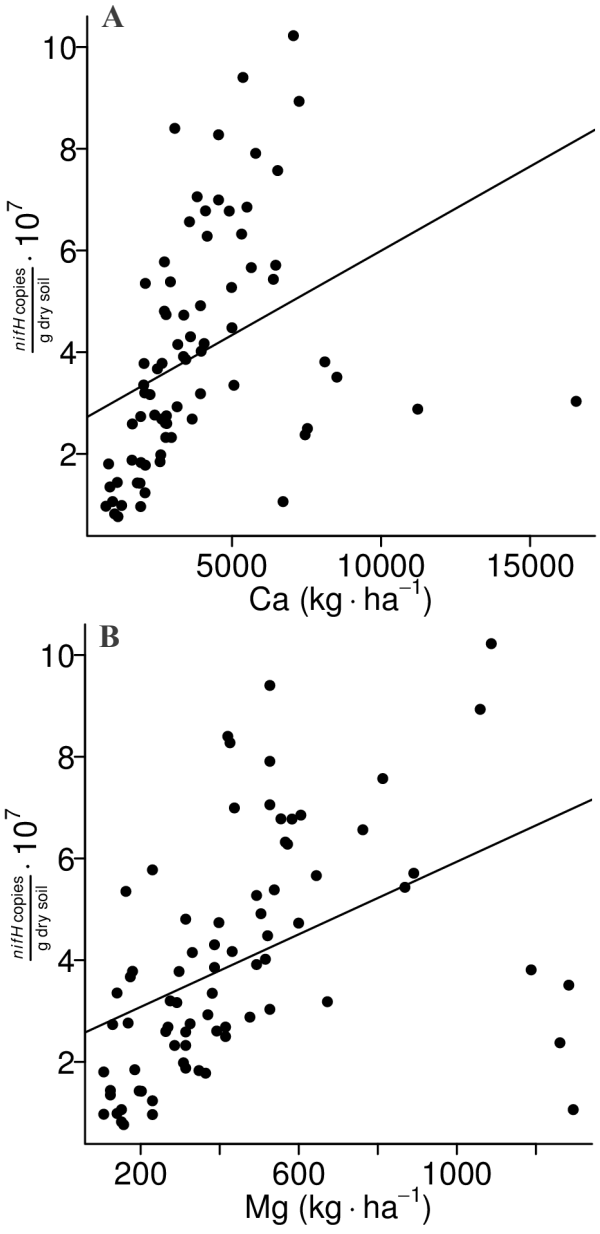


Figure 6.9 The positive relationship between soil moisture as percent water (Pw) and relative *nifH* abundance. Low soil moisture constrains nitrogen fixer abundance. The $R^2 = 0.2077$ and $p = 2.367 \times 10^{-5}$ for the association. One univariate Pw outlier was removed and the R^2 with the outlier included is 0.1983. (n=75).

fixation are generally lower than those of tropical savanna and tropical forests by about an order of magnitude, though in both these cases a larger contribution is estimated to derive from symbiotic organisms [1]. It should also be kept in mind though that our laboratory assay, which was intended to compare potential nitrogen fixation rates across various sites, does not mimic the natural environment because the plant roots have been removed and the original soil structure was disturbed by sieving. Also, the time of sampling may have affected the rate of nitrogen fixation given that in

Figure 6.10 A strong positive relationship exists between soil Ca (A) and Mg (B) with soil-normalized nifH abundance. For the association with Ca, $R^2 = 0.1246$ and $p = 0.001032$; with Mg the $R^2 = 0.1902$ and $p = 4.4885 \times 10^{-5}$.


A
10
8
6
4
2
$\frac{nifH\ copies}{g\ dry\ soil} \cdot 10^7$
5000
10000
15000
Ca (kg · ha^{-1})
B
10
8
6
4
2
$\frac{nifH\ copies}{g\ dry\ soil} \cdot 10^7$
200
600
1000
Mg (kg · ha^{-1})

the fall the plants would have been senescing and would be inferred to have lower rates of rhizodeposition because of decreased productivity. The abundance of nitrogen fixers has been noted to be higher in the rhizosphere as compared to bulk soil [13], and this occurs as a result of labile C substrates exuded by roots which promote nitrogen fixation [36]. Theoretical estimation of the range of nitrogen fixation that can take place as a result of rhizodeposition was determined to vary from 0.2 to 4 kg N ha^{-1} yr^{-1} [37]. Thus, our rates fall within range of expected values.

By exploring the relationships between the environmental variables and our measured response variables which were nitrogen fixation rate and nitrogen fixer abundance, we found that most sites fell into one of two δ ^{13}C isotopic signature groups (Figure 6.1A). Upon further exploration, we saw that these two groups also have distinct δ ^{15}N signatures as well (Figure 6.1B), and taken together these indicate distinct isotopic groups (Figure 6.2) that are likely to be undergoing unique soil nitrogen- and carbon-cycling processes. Characterization of plant δ ^{13}C has revealed that plant taxa exhibit distinct isotopic signatures in their biomass, with the greatest separation between plants which carry out C3 and C4 photosynthesis [38]. C3 plants have an average δ ^{13}C of -27 ‰ whereas C4 plants have one of -12 ‰ [38]. Thus, one explanation for the bimodal δ ^{13}C distribution could be the dominance of two distinct C3 plant taxa which differ in the δ ^{13}C signature of their biomass. Old field sites in this region are generally vegetated by goldenrod (*Solidago* species), Aster (*Aster* species), and grasses [39], and some sites like AW and AN were observed to be dominated by

goldenrod at the time of sampling. The depleted mode of δ ^{13}C values (Figure 6.1A) corresponds to a value measured for *Solidago missouriensis* biomass at -27.4 ‰ [40], which suggests that *Solidago* may be the biomass origin for the depleted δ ^{13}C group, though we did not identify the particular *Solidago* species which occurred in our sample plots, and we did not survey the isotopic signature of biomass for different species in our sampling sites. An alternative explanation is that the bimodal nature of the δ ^{13}C values is a result of higher rates of a microbial process in one of the δ ^{13}C groups, and the separation we see between the two modes is consistent with an observation that microbial biomass has a δ ^{13}C isotopic signature that is 1.6 ‰ enriched relative to total soil C for soils under C3 vegetation [41].

Our results show increased nitrogen mineralization in the enriched δ ^{13}C group

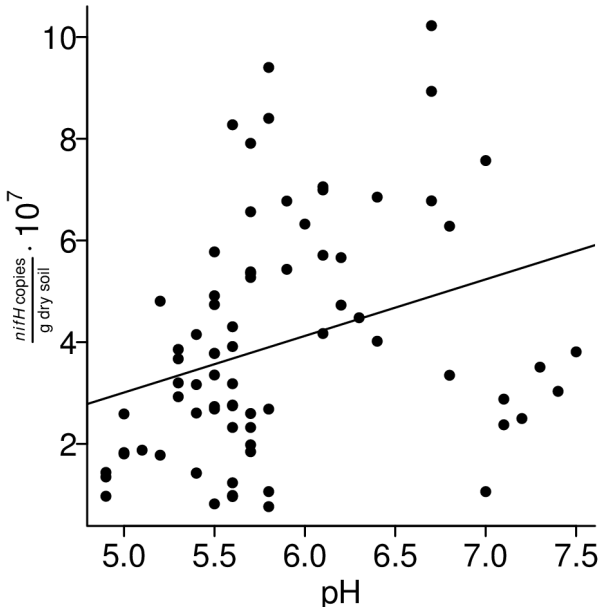


Figure 6.11 The positive correlation of pH and absolute *nifH* abundance. R^2 = 0.08363, p=0.006498.

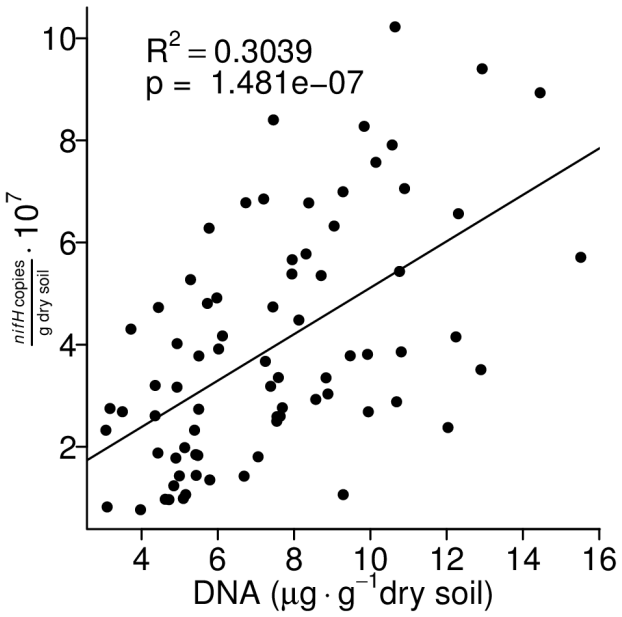


Figure 6.12 The positive correlation of extracted soil DNA and absolute *nifH* abundance.

and evidence for nitrogen fixation in the depleted δ ^{13}C group. The depleted δ ^{13}C group partitioned closer to 0 in the δ ^{15}N values (Figure 6.1B) suggesting this group is actively fixing nitrogen. Nitrogen fixation rates are different between the groups (Figure 6.3A), and a number of low values occur in the depleted group (Figure 6.3A) suggesting that nitrogen fixation is constrained in many samples under our assay conditions. We see that the depleted δ ^{13}C group also has a higher relative abundance of *nifH* (Figure 6.6), which could indicate selective advantage of nitrogen fixers on what could be biomass provided by distinct plant taxa. Alternatively, this may indicate an increased abundance of certain nitrogen-fixing bacteria in association with the depleted group. In contrast, the nitrogen mineralization rate correlates with δ ^{15}N, with

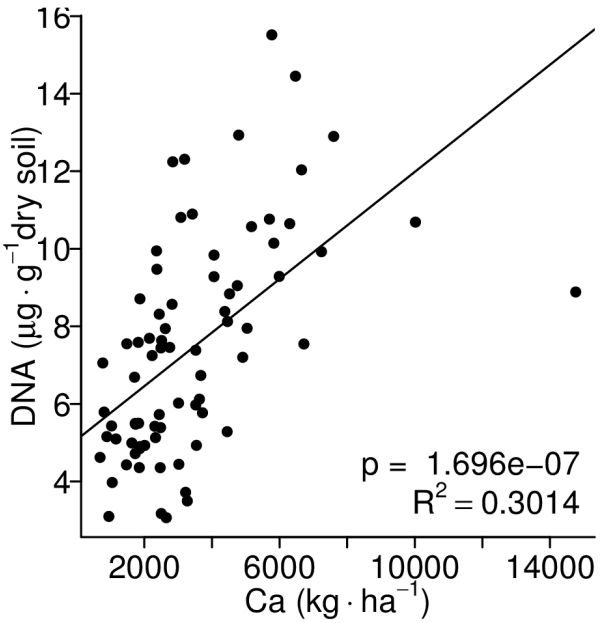


Figure 6.13 The positive correlation of soil Ca and extracted soil DNA.

higher rates occurring among enriched δ ^{15}N values (Figure 6.4A). Nitrogen mineralization is known to result in enrichment of ^{15}N in microbial biomass [47], and this corresponds to the higher rates of nitrogen mineralization observed among samples with higher δ ^{15}N values (Figure 6.4A). We do not witness co-occurrence of high nitrogen mineralization with nitrogen fixation (Figure 6.5). Also, high rates of nitrogen mineralization do not co-occur with high *nifH* relative abundance (Figure 6.7). This is consistent with the occurrence of nitrogen fixation, which requires expenditure of energy to occur, when other sources are not available. Further supporting that nitrogen mineralization is the dominant process in the δ ^{13}C enriched group is evidence which comes from the soil K and P (Figure 6.8 A and B). The presence of long-term recycling of plant organic matter would drive up the

concentration of these biomass-associated variables and provide the organic matter to fuel the increased nitrogen mineralization we see (Figure 6.4A). However, organic matter and soil moisture were correlate in our data, and the nitrogen mineralization rate could be a response to differences in soil moisture that occur in the sites throughout the year. Furthermore, the enriched group shows overall low *nifH* relative abundance which is to be expected if the dominant process is nitrogen mineralization. Thus, these two processes appear to be incompatible, with selective pressure against nitrogen fixers in soils where nitrogen mineralization predominates.

It is important to note that normalization of the qPCR data by two common approaches affected a number of the associations we observed between *nifH* copy number and environmental variables. We normalized our qPCR data to the dry soil weight and also to the amount of DNA recovered from the soil extraction. If the amount of DNA recovered from the soil is taken as a representation of the overall microbial community size, then *nifH* copies μg^{-1} soil DNA would represent the proportion of the microbial community that are nitrogen fixers targeted by the PolF/PolR primer set, and increases in this measure of relative abundance would indicate selective pressure in favor of nitrogen fixation, and is perhaps a stronger indicator of what conditions are selective for nitrogen fixers. On the other hand, *nifH* copies g^{-1} dry soil indicates the absolute abundance of the nitrogen fixers. Increases in this absolute measure could point to an overall higher capacity for nitrogen fixation, but alternatively, if the presence of the *nifH* gene is taken to occur in a fixed

percentage of the community members regardless of whether it is actively used, then this may track increases in the overall abundance of the bacterial community, thus limiting its utility to probe the drivers of nitrogen fixation.

The absolute abundance of *nifH* was correlated to soil moisture as well as pH and its related variables Ca and Mg. Low abundances of *nifH* occurred at low soil moisture (Figure 6.9) which meets expectation since moisture is a general biological control for the activity of microbial processes [42, 43]. The variables Ca and Mg were strongly correlated with *nifH* abundance (Figure 6.10 A and B), as was pH but to a lesser extent (Figure 6.11). Correlations of pH with *nifH* have been observed in other studies [8] as well as with overall bacterial abundance [44]. The absolute abundance of *nifH* was correlated to the amount of DNA extracted (Figure 6.12), and the amount of Ca was correlated to the amount of DNA extracted. This latter correlation is a concern because, as has been well-demonstrated before, clays in soil bind more DNA as soil Ca increases [50]. The mechanism for this increase involves the formation of a cationic bridge by Ca between the negatively charged clay and the negatively charged phosphate groups on the DNA. Once adsorbed, the DNA is protected from degradation [51], and would thus have a longer residence time in the soil leading to higher DNA recovery upon extraction. That *nifH* abundance has been observed to increase with pH may not be revealing pH as a driver of nitrogen fixer abundance, but rather it describes the chemistry of DNA adsorption onto clays and subsequent protection of DNA from degradation as Ca increases in the soil.

Correlations were not observed with some variables that would have been hypothesized to correlate with either nitrogen fixation rate or universal *nifH* abundance. Nitrogen fixation did not correlate with pH, nor with *nifH* abundance, nor was there a regional difference in the nitrogen fixation rates measured ($p<0.05$). We measured endogenous rates of nitrogen fixation reflective of the conditions present in the soil at the time of sampling, and did not measure potential rates. The C to N ratio with a mean of 11.5 was uninformative, though the range of values over which it varied was closer to the 8:1 ratio of bacterial biomass than to that of plant biomass known to stimulate nitrogen fixation [45]. When bivariate comparisons were made of the nitrogen fixation rates for all 20 sites with variables like C:N, *nifH* copy number, pH, and other variables with which nitrogen fixation may be reasonably hypothesized to correlate, no correlations were witnessed. The absence of a relationship may stem from limitations in the laboratory-based assay, or could be due to numerous environmental constraints on nitrogen-fixation like biomass availability, pH, or soil moisture. Our laboratory assay for potential nitrogen fixation rate did not equilibrate the soil water-filled pore space as was done for the nitrogen mineralization assay, and making such an adjustment could give more consistent results given that water-filled pore space is known to affect microbial processes [43].

By sampling 20 old field sites in two geographic regions of New York State with similar soil textures, recent land use, and vegetation cover, we have shown that nitrogen fixation and nitrogen mineralization are incompatible processes whereby one

process or the other predominates and distinguishes sites. Isotopic signatures for C and N point to higher nitrogen fixation rates and greater abundance of nitrogen fixers in old fields which are isotopically depleted whereas old fields enriched by 1.25 ‰ in $\delta^{13}C$ are dominated by nitrogen mineralization. We also identified pH as a variable which correlates with the absolute abundance of nitrogen fixers but this may be a result of a correlation of DNA recovery to pH. Low soil moisture was observed to act as a general biological constraint which limits nitrogen fixer abundance. Future work should determine whether the separation in $\delta^{13}C$ signatures of the old fields is due to differences in the dominant plant species serving as biomass inputs to the soil or due to high rates of microbial processes like nitrogen mineralization.

References

1. Cleveland C, Townsend A, Schimel D, Fisher H, Howarth R et al. (1999) Global patterns of terrestrial biological nitrogen (N_2) fixation in natural ecosystems. Global Biogeochem Cycles 13: 645.

2. Gaby JC, Buckley DH (2011) A global census of nitrogenase diversity. Environ Microbiol 13: 1790-1799.

3. Rappe M, Giovannoni S (2003) The uncultured microbial majority. Annu Rev Microbiol 57: 394.

4. Zehr JP, Jenkins BD, Short SM, Steward GF (2003) Nitrogenase gene diversity and microbial community structure: a cross-system comparison. Environ Microbiol 5: 539-554.

5. Igarashi RY, Seefeldt LC (2003) Nitrogen fixation: the mechanism of the Mo-dependent nitrogenase. Crit Rev Biochem Mol Biol 38: 351-384.

6. Smith CJ, Osborn AM (2009) Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. FEMS Microbiol Ecol 67: 6-20.

7. Hayden HL, Drake J, Imhof M, Oxley APA, Norng S et al. (2010) The abundance of nitrogen cycle genes *amoA* and *nifH* depends on land-uses and soil types in South-Eastern Australia. Soil Biol Biochem 42: 1774-1783.

8. Pereira e Silva MC, Semenov AV, van Elsas JD, Salles JF (2011) Seasonal variations in the diversity and abundance of diazotrophic communities across soils. FEMS Microbiol Ecol 77: 57-68.

9. Wakelin SA, Gregg AL, Simpson RJ, Li GD, Riley IT et al. (2009) Pasture management clearly affects soil microbial community structure and N-cycling bacteria. Pedobiologia 52: 237-251.

10. Juraeva D, George E, Davranov K, Ruppel S (2006) Detection and quantification of the *nifH* gene in shoot and root of cucumber plants. Can J Microbiol 52: 731-739.

11. Hai B, Diallo NH, Sall S, Haesler F, Schauss K et al. (2009) Quantification of key genes steering the microbial nitrogen cycle in the rhizosphere of sorghum cultivars in tropical agroecosystems. Appl Environ Microbiol 75: 4993-5000.

12. Babić KH, Schauss K, Hai B, Sikora S, Redzepović S et al. (2008) Influence of

different Sinorhizobium meliloti inocula on abundance of genes involved in nitrogen transformations in the rhizosphere of alfalfa (Medicago sativa L.). Environ Microbiol 10: 2922-2930.

13. Coelho MRR, Marriel IE, Jenkins SN, Lanyon CV, Seldin L et al. (2009) Molecular detection and quantification of *nifH* gene sequences in the rhizosphere of sorghum (*Sorghum bicolor*) sown with two levels of nitrogen fertilizer. Applied Soil Ecology 42: .

14. Morales SE, Cosart T, Holben WE (2010) Bacterial gene abundances as indicators of greenhouse gas emission in soils. ISME J 4: 799-808.

15. Martensson L, Diez B, Wartiainen I, Zheng W, El-Shehawy R et al. (2009) Diazotrophic diversity, *nifH* gene expression and nitrogenase activity in a rice paddy field in Fujian, China. Plant and Soil 325: 207-218.

16. Kroeckel L, Stolp H (1984) Influence of soil-water potential on respiration and nitrogen-fixation of azotobacter-vinelandii. Plant Soil 79: 37-49.

17. Kapustka LA, Rice ER (1978) Acetylene reduction nitrogen fixation of glucose amended soils from central oklahoma usa old field succession plots. Southwestern Naturalist 23: 389-396.

18. Hegazi N, Vlassak K, Monib M (1979) Effect of amendments, moisture and temperature on acetylene-reduction in nile delta soils. Plant Soil 51: 37.

19. O'toole P, Knowles R (1973) Efficiency of acetylene reduction nitrogen fixation in soil effect of type and concentration of available carbohydrate. Soil Biol Biochem 5: 797.

20. Roper MM (1985) Straw decomposition and nitrogenase activity (C_2H_2 reduction) - effects of soil-moisture and temperature. Soil Biol Biochem 17: 65-71.

21. Barrow NJ, Jenkinson DS (1962) The effect of water-logging on fixation of nitrogen by soil incubated with straw. Plant Soil 16: 258-262.

22. Roper MM (1983) Field-measurements of nitrogenase activity in soils amended with wheat straw. Aust J Ag Res 34: 739.

23. Krotzky A, Berggold R, Werner D (1986) Analysis of factors limiting associative N_2-fixation (C_2H_2 reduction) with 2 cultivars of sorghum-nutans. Soil Biol Biochem 18: 207.

24. Weber A, Niemi M, Sundman V, Skujins J (1983) Acetylene-reduction (N_2 fixation) and endogenous ethylene release in sub-boreal soils and peats of finland. Oikos 41: 226.

25. Roper M, Smith N (1991) Straw decomposition and nitrogenase activity (C_2H_2 reduction) by free-living microorganisms from soil - effects of pH and clay content. Soil Biol Biochem 23: 283.

26. Vitousek PM, Cassman K, Cleveland C, Crews T, Field CB et al. (2002) Towards an ecological understanding of biological nitrogen fixation. Biogeochemistry 57: 1-45.

27. Elliott ET, Heil JW, Kelly EF, Monger HC (1999) Soil structural and other physical properties. In: Robertson G, Coleman D, Bledsoe C, Sollins P, editors. Standard soil methods for long-term ecological research. Oxford University Press, USA. pp. 74-85.

28. Various (2004) Soil survey laboratory methods manual. USDA NRC.

29. Various (1995) Recommended soil testing procedures for the northeastern united states. Northeast Coordinating Committee on Soil Testing (NEC-67) - College of Agriculture and Natural Resources - University of Delaware: Newark, US (DE).

30. Various (1982) Methods of soil analysis. Page, AL editor. American Society of Agronomy.

31. Robertson GP, Wedin D, Groffman PM, Blair JM, Holland EA, Nedelhoffer KJ, Harris D (1999) Soil carbon and nitrogen availability. nitrogen mineralization, nitrification and soil respiration potentials. In: Robertson G, Coleman D, Bledsoe C, Sollins P, editors. Standard soil methods for long-term ecological research. Oxford University Press, USA. pp. 258-271.

32. Poly F, Monrozier LJ, Bally R (2001) Improvement in the RFLP procedure for studying the diversity of *nifH* genes in communities of nitrogen fixers in soil. Res Microbiol 152: 95-103.

33. Grubbs FE (1950) Sample criteria for outlying observations. Annals of Mathematical Statistics 21: 27-58.

34. Belnap J (2001) Factors influencing nitrogen fixation and nitrogen release in biological soil crusts. Ecological Studies. Biological soil crusts: Structure, function, and management 150: 241-261.

35. Wurzburger N, Bellenger JP, Kraepiel AML, Hedin LO (2012) Molybdenum and phosphorus interact to constrain asymbiotic nitrogen fixation in tropical forests. PloS one 7: e33710-e33710.

36. Burgmann H, Meier S, Bunge M, Widmer F, Zeyer J (2005) Effects of model root exudates on structure and activity of a soil diazotroph community. Environ Microbiol 7: 1724.

37. Jones DL, Farrar J, Giller KE (2003) Associative nitrogen fixation and root exudation - What is theoretically possible in the rhizosphere? Sybiosis 35: 19-38.

38. Staddon PL (2004) Carbon isotopes in functional soil ecology. Trends Ecol Evol 19: 148-154.

39. Mohler, Charles L,, Marks PL, Gardescu S (2006) Guide to the plant communities of the central finger lakes region. N.Y.S. Agricultural Experiment Station, Communications Services, Geneva, N.Y.

40. Still CJ, Berry JA, Ribas-Carbo M, Helliker BR (2003) The contribution of C3 and C4 plants to the carbon cycle of a tallgrass prairie: an isotopic approach. Oecologia 136: 347-359.

41. Dijkstra P, Ishizu A, Doucett R, Hart SC, Schwartz E et al. (2006) C-13 and N-15 natural abundance of the soil microbial biomass. Soil Biol Biochem 38: Amer Geophys Union-3266.

42. Linn D, Doran J (1984) Effect of water-filled pore-space on carbon-dioxide and nitrous-oxide production in tilled and nontilled soils. Soil Sci Soc Am J 48: 1272.

43. Franzluebbers AJ (1999) Microbial activity in response to water-filled pore space of variably eroded southern Piedmont soils. Appl Soil Ecol 11: 91-101.

44. Rousk J, Bååth E, Brookes PC, Lauber CL, Lozupone C et al. (2010) Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME J 4: 1340-1351.

45. Perez CA, Carmona MR, Armesto JJ (2010) Non-symbiotic nitrogen fixation during leaf litter decomposition in an old-growth temperate rain forest of Chiloe Island, southern Chile: effects of single versus mixed species litter. Austral Ecology 35: 148-156.

46. Kapustka LA, Rice, EL (1978) Acetylene reduction N_2-fixation of glucose-

amended soils from central Oklahoma old field succession plots. The Southwestern Naturalist 23(3):389-396.

47. Dijkstra,P. et al. (2008) ^{15}N enrichment as an integrator of the effects of C and N on microbial metabolism and ecosystem function. Ecology letters 11: 389–97.

48. Roskoski, JP (1980) Nitrogen fixation in hardwood forests of the northeastern United States. Plant and Soil 54:33-44.

49. Reed, SC, Cleveland, CC, and Townsend, AR (2008) Tree species control rates of free-living nitrogen fixation in a tropical rain forest. Ecology 89(10)2924-2934.

50. Poly, F. et al. (2000) Differences between linear chromosomal and supercoiled plasmid DNA in their mechanisms and extent of adsorption on clay minerals. Langmuir 16: 1233–1238.

51. Demanèche, S. et al. (2001) Evaluation of biological and physical protection against nuclease degradation of clay-bound plasmid DNA. Appl Environ Microbiol 67: 293–299.

CHAPTER 7

FUTURE DIRECTIONS

Inevitably, once research has been summarized in writing and submitted as a publication, new ideas will have arisen which, whether for lack of time or resources, cannot be realized. The following is a summary of my thoughts on where this work could be taken next.

Chapter 2

The *nifH* database requires substantial effort to update. Currently, the Zehr group maintains a similar database. It would save time and effort to share responsibility for maintaining this community resource among the users. Maintenance effort may be reduced by scripting the update and alignment process. Additionally, there are a number of *nifD* and *nifK* sequence fragments from PCR-based studies which may be brought into the database.

With regards to the evolutionary analysis, more genes, as both controls and further genes from the *nif* operon, may be brought into the database from sequenced genomes for additional analysis.

Chapter 3

The diversity analyses should be repeated after the amount of sequence

information has expanded. Work in chapter 2, though it occurred chronologically later than the work in chapter 3, informs the choice of OTU cutoff, showing that *nifH* may vary by as much as 20% without exceeding 97% similarity in the 16S gene. Thus, higher OTU cutoffs may be justified for assessing *nifH* diversity, though attempts at identifying species by use of *nifH* is advised against.

Chapter 4

The scripts written for assessing *nifH* primer coverage should be rewritten in Perl, Python, or another suitable language to serve as a stand-alone tool for processing a sequence alignment and a list of primers to generate a table of primer coverage. This would allow other researchers to readily assess a set of primers against all available sequence information for their gene, so long as a reliable alignment may be constructed. Also, the primers are assessed by simple text-matching, and an algorithm could be written that takes account of the fact that a mismatch at the 5' end of a primer will not reduce priming as would a mismatch at the 3' end. This would require empirical or modeling work to be able to quantify the effects on priming efficiency for positions along primers of varying length and nucleotide composition.

Chapter 5

The template-specific bias summarized in this chapter has now been described elsewhere. Specific experiments may be devised to determine the aspects of a primer that predispose it to bias.

Chapter 6

The latitude/longitude coordinates for each sampling transect in the old field survey is recorded down to several meters accuracy. Thus, follow-up samples may be taken at nearly the exact location where the first samples where obtained. In order to determine the factor which lead to two groups with distinct isotopic values among the sites, the most straightforward approach would be to sample vegetation for isotopic signature to see if certain plants correspond to the one of the isotopic groups. The stable isotope assay for determining nitrogen fixation rate should begin by equilibrating the water-filled pore space for each sample to an optimum, which may be determined experimentally. Microcosm studies may begin to pull apart the many constraints on nitrogen fixation, whereby individual factors like litter type, pH, and other variables may be varied individually for soils from contrasting sites.